

# A Transparency Index Framework for Machine Learning powered AI in Education

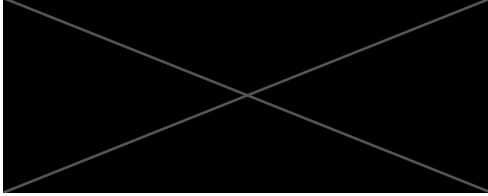
Muhammad Ali Chaudhry

Institute of Education, University College London

Submitted for the Degree of Doctor of Philosophy

## **Author's Declaration**

I, Muhammad Ali Chaudhry, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.



Date of Submission: 5<sup>th</sup> June 2022

## Abstract

The increase in the use of AI systems in our daily lives, brings calls for more ethical AI development from different sectors including, finance, the judiciary and to an increasing extent education. A number of AI ethics checklists and frameworks have been proposed focusing on different dimensions of ethical AI, such as fairness, explainability and safety. However, the abstract nature of these existing ethical AI guidelines often makes them difficult to operationalise in real-world contexts. The inadequacy of the existing situation with respect to ethical guidance is further complicated by the paucity of work to develop *transparent* machine learning powered AI systems for real-world. This is particularly true for AI applied in education and training.

In this thesis, a Transparency Index Framework is presented as a tool to forefront the importance of transparency and aid the contextualisation of ethical guidance for the education and training sector. The transparency index framework presented here has been developed in three iterative phases.

In phase one, an extensive literature review of the real-world AI development pipelines was conducted. In phase two, an AI-powered tool for use in an educational and training setting was developed. The initial version of the Transparency Index Framework was prepared after phase two. And in phase three, a revised version of the Transparency Index Framework was co-designed that integrates learning from phases one and two. The co-design process engaged a range of different AI in education stakeholders, including educators, ed-tech experts and AI practitioners.

The Transparency Index Framework presented in this thesis maps the requirements of transparency for different categories of AI in education stakeholders, and shows how transparency considerations can be ingrained throughout the AI development process, from initial data collection to deployment in the world, including continuing iterative improvements. Transparency is shown to enable the implementation of other ethical AI dimensions, such as interpretability, accountability and safety. The

optimisation of transparency from the perspective of end-users and ed-tech companies who are developing AI systems is discussed and the importance of conceptualising transparency in developing AI powered ed-tech products is highlighted. In particular, the potential for transparency to bridge the gap between the machine learning and learning science communities is noted. For example, through the use of datasheets, model cards and factsheets adapted and contextualised for education through a range of stakeholder perspectives, including educators, ed-tech experts and AI practitioners.

## Impact Statement

Mishaps with AI in the world beyond the lab are not new. Various stakeholders in the financial services, healthcare, recruitment, law enforcement and e-commerce sectors have suffered due to AI not performing as expected in different contexts. UK's A-level grading fiasco highlights the impact of AI going wrong within education. The Transparency Index Framework for AI-powered ed-tech products developed in this research aims to avoid such incidents and has both academic and non-academic impact.

Academically, it is one of the first research-based frameworks focusing on the transparency of AI-powered ed-tech products. It can potentially help learning scientists and educational researchers to

- Get a better understanding of how an AI-powered ed-tech has been built;
- Reproduce and build on the research and results claimed by an AI-powered ed-tech;
- Have a detailed overview of the deficiencies of the data used, assumptions made and decisions taken during the AI development process.

In academia, the Transparency Index Framework can help in the identification of the gap between researchers in education-related social science domains (such as learning sciences, cognitive sciences, early childhood development, lifelong learning, primary, secondary and higher education) and more technical domains like machine learning and AI research. It potentially offers social science academics a checklist and an auditing framework to focus on, when evaluating AI-powered products. For AI researchers and engineers focusing on the development of AI-powered products, the Transparency Index offers a detailed documentation framework that enables robust development and deployment of AI-powered ed-tech in real-world.

For ed-tech companies who plan to develop AI-powered ed-tech, the Transparency Index Framework can be utilised to provide detailed guidelines

to ensure ethical AI development. It covers the entire AI development pipeline, from brainstorming ideas and data collection to the deployment and iterative improvements of AI in the real-world. The Transparency Index Framework is not only useful for ed-tech companies to build safe and robust AI products, but it also helps:

- AI practitioners to justify the assumptions and document the decisions taken during the AI development process;
- Educators to formulate the kind of questions they need to ask ed-tech companies before deploying AI-powered products in their institutions;
- Ed-tech experts to evaluate the pros and cons of AI-powered ed-tech;
- Regulators to audit AI-powered ed-tech, identify any deficiencies in the development process and hold relevant personnel accountable if anything goes wrong.

In the past few years there has been a significant shift to online learning. It means ed-tech companies have more data than ever before to build AI-powered products and generate more business value. In this context, the development of the Transparency Index Framework for AI-powered ed-tech companies, educators, ed-tech experts and AI practitioners is most timely. It can help AI in education stakeholders to build and use safe and secure AI systems within the contexts in which they are expected to perform optimally.

## **Dedication**

We are a product of the greatest influences of our lives. For me, my family have been the biggest influence in my life, and I dedicate this work to them.

To my father, who believed in me and instilled confidence in me whenever I was going through the toughest times. Who never uttered a single word of discouragement or disappointment whenever I failed, and whose belief in me has always encouraged me to take risks and make bold decisions. I wish I could treat my kids with similar optimism and positivity.

To my mother, whose empathy and generosity towards the destitute inspired me to initiate Ali Foundation (previously Renaissance) and institutionalize our philanthropic activities. To my elder brother (Umer) and sisters (Rabia and Hajra) who have always cared for me and tolerated me through highs and lows.

To my wife Zainab, the love of my life; who has always stood by me and shown me light through the darkest tunnels. I could not have asked for a better partner.

Last, but not the least, I dedicate this work to the children studying in Ali Foundation's schools across Rawalpindi and Islamabad. The hard work and struggle these children face from childhood is truly inspirational and I wish to see at least one of our students complete a PhD as well.

## **Acknowledgements**

I believe everyone in this world is born lucky in some respect. I consider myself extremely lucky to have Prof Rose Luckin and Associate Prof Mutlu Cukurova as my supervisors, my teachers, my mentors and my inspiration.

Rose – I will always be indebted to you for the support and guidance you provided me during the PhD while I moved to UK, got married, had two kids and moved back due to the pandemic - thank you very much for your enduring conviction and belief. Your humility is inspirational for early career researchers like me, and you are truly a blessing for the AIED community. Proud to be your student, forever.

Mutlu – I would not have done this without you. Thank you for your patience, your positive and healthy criticism, your infectious drive, your undying energy, your exemplary hard work, and most importantly your encouragement throughout. I will always be deeply grateful.



# Table of Contents

<b>List of Figures .....</b>	<b>13</b>
<b>List of Tables .....</b>	<b>14</b>
<b>Chapter 1: Background and Research Issue .....</b>	<b>16</b>
<b>1.1 Introduction .....</b>	<b>16</b>
<b>1.2 Ethical AI .....</b>	<b>18</b>
<b>1.3 Bias .....</b>	<b>23</b>
<b>1.4 Why Transparency? .....</b>	<b>28</b>
<b>1.5 Research Questions: .....</b>	<b>31</b>
<b>1.6 Methods.....</b>	<b>34</b>
1.6.1 Framework Creation Methodology.....	35
1.6.1.1 Ethical Considerations .....	35
1.6.1.2 Phase 1 .....	36
1.6.1.3 Phase 2 .....	37
1.6.2 Framework Evaluation Methodology.....	40
<b>1.7 My role in the project .....</b>	<b>40</b>
<b>1.8 Conclusion .....</b>	<b>41</b>
<b>1.9 Glossary.....</b>	<b>42</b>
<b>Chapter 2: Phase 1 Research Background: Transparency, Artificial Intelligence and Education .....</b>	<b>45</b>
<b>2.1 Introduction .....</b>	<b>45</b>
<b>2.2 The variability of a ‘Process’ .....</b>	<b>46</b>
<b>2.3 The variety of ‘Information’ .....</b>	<b>48</b>
<b>2.4 The interpretations of ‘Shared’ .....</b>	<b>48</b>
<b>2.5 The variety of meanings of ‘Enhancing the Understanding’ .....</b>	<b>50</b>
<b>2.6 Transparency and other ethical AI dimensions.....</b>	<b>51</b>
2.6.1 Explainability and Interpretability .....	53
2.6.2 Fairness .....	54
2.6.3 Accountability .....	56
2.6.4 Safety .....	58
<b>2.7 The Artificial Intelligence Development Pipeline.....</b>	<b>62</b>
<b>2.8 An AI Development Pipeline in Education .....</b>	<b>63</b>
2.8.1 Planning.....	64
2.8.2 Data Collection and Engineering .....	67
2.8.3 Machine Learning Modelling.....	71
2.8.4 Deployment and Improvements .....	77

2.8.4.1 Facial Recognition .....	79
<b>2.9 Gaps in the Literature .....</b>	<b>81</b>
<b>2.10 AI-powered Ed-tech tool .....</b>	<b>82</b>
<b>2.11 Conclusion .....</b>	<b>83</b>
<b>Chapter 3: Phase 2 - Framework Creation: Data Processing Stage .....</b>	<b>85</b>
<b>3.1 Introduction .....</b>	<b>85</b>
<b>3.2 Documenting the Dataset .....</b>	<b>89</b>
3.2.1 Datasheet for the AI-powered Tool .....	91
3.2.1.1 Motivation for Dataset Creation: .....	91
3.2.1.2 Dataset Composition .....	92
3.2.1.3 Data Collection Process .....	96
3.2.1.4 Data Pre-processing: .....	100
3.2.1.5 Dataset Maintenance .....	102
3.2.1.6 Legal & Ethical Considerations: .....	102
<b>3.3 Testing the Assumptions .....</b>	<b>106</b>
3.3.1 Nine Month Analysis Report .....	107
3.3.1.1 Monthly Traders' Performance .....	107
3.3.1.2 Exploratory Analysis .....	108
3.3.1.3 Statistical Analysis .....	111
3.3.1.4 Auto Correlation Plots .....	112
3.3.1.5 Consecutive Month Differences .....	115
3.3.1.6 Summary Statistics .....	118
<b>3.4 Conclusion .....</b>	<b>119</b>
<b>Chapter 4: Phase 2 - Framework Creation: Machine Learning Modelling Stage .....</b>	<b>122</b>
<b>4.1 Introduction .....</b>	<b>122</b>
<b>4.2 Models Evaluation Report .....</b>	<b>124</b>
4.2.1 Introduction .....	124
4.2.2 Machine Learning Models .....	125
4.2.3 Accuracy Measures .....	128
4.2.4 Ethical Considerations for Model Selection .....	129
4.2.5 Main Results .....	130
4.2.6 Conclusion .....	138
<b>4.3 Model Card .....</b>	<b>140</b>
4.3.1 Model Details .....	141
4.3.2 Intended Use: .....	142
4.3.3 Factors: .....	142
4.3.4 Metrics: .....	143

4.3.5 Training Data .....	143
4.3.6 Quantitative Analysis .....	144
4.3.7 Ethical Considerations (as specified in the Model Card) .....	144
4.3.8 Caveats and Recommendations .....	145
<b>4.4 Conclusion .....</b>	<b>146</b>
<b>Chapter 5: Phase 2 - Framework Creation: Deployment and Iterative Improvements Stage .....</b>	<b>148</b>
<b>5.1 Introduction .....</b>	<b>148</b>
<b>5.2 Model Validation Report .....</b>	<b>150</b>
5.2.1 Introduction .....	150
5.2.2 Configurations of the tool .....	151
5.2.3 Analysis .....	153
5.2.3.1 Indian Office .....	157
5.2.3.2 Polish Office.....	158
5.2.4 Discussion.....	161
5.2.5 Limitations and future investigations.....	163
<b>5.3 Conclusion .....</b>	<b>164</b>
5.3.1 First version of the Transparency Index Framework.....	165
<b>Chapter 6: Phase 3 - Framework Creation Part 4: Evaluating and Improving the Transparency Index Framework.....</b>	<b>170</b>
<b>6.1 Introduction .....</b>	<b>170</b>
<b>6.2. Framework Evaluation.....</b>	<b>172</b>
<b>6.3 Themes .....</b>	<b>175</b>
6.3.1 The usefulness of the proposed framework .....	175
6.3.2 Transparency of AI products “a new phenomenon” .....	176
6.3.3 Focus on Transparency and Ethics in AI products .....	176
6.3.4 Feedback and Recommendations to Improve the Framework .....	179
<b>6.4 Conclusion .....</b>	<b>180</b>
<b>Chapter 7: Phase 3 - The Revised Transparency Index Framework .....</b>	<b>182</b>
<b>7.1 Introduction .....</b>	<b>182</b>
<b>7.2 Framework Description .....</b>	<b>182</b>
<b>7.3 Transparency Index Framework .....</b>	<b>183</b>
<b>7.4 Discussion: Transparency in the context of AI in Education .....</b>	<b>191</b>
<b>7.5 Conclusion .....</b>	<b>203</b>
<b>Chapter 8: Conclusion and Future Work .....</b>	<b>205</b>
<b>8.1 Limitations .....</b>	<b>207</b>
<b>8.2 Contribution.....</b>	<b>208</b>

<b>8.3 Future Work.....</b>	<b>209</b>
<b><i>References.....</i></b>	<b>211</b>
<b><i>Appendices.....</i></b>	<b>251</b>
<b>Appendix 1: The details of different features used in personality surveys .....</b>	<b>251</b>
<b>Appendix 2: The Survey’s English version of the reduced set of items .....</b>	<b>266</b>
<b>Appendix 3: Distribution of scores for each Feature used in Training Data .....</b>	<b>292</b>
<b>Appendix 4: Informed consent form used in phase 2 .....</b>	<b>295</b>
<b>Appendix 5: Informed consent form used in phase 3 .....</b>	<b>296</b>
<b>Appendix 6: Questions for Interviews in phase 3 .....</b>	<b>297</b>
Stage 1 Questions: .....	297
Stage 2 Questions: .....	298
<b>Appendix 7: Ethics Approval from the UCL Research Ethics Committee.....</b>	<b>300</b>

## List of Figures

Figure 1: Search for Common Cognitive Ground .....	22
Figure 2: Steps involved in the development of an AI product with different types of biases that can occur in different stages .....	24
Figure 3a: Research question and contributions of this research.....	33
Figure 4: Potential overlap between Transparency and other ethical AI dimensions	52
Figure 5: Building Safe Artificial Intelligence: Specification, Robustness and Assurance .....	59
Figure 6: The importance of transparency in bridging the gap between Ideal and Revealed specifications of an AI system .....	61
Figure 7: The canonical architecture for AI tool development in terms of transparency .....	64
Figure 8: An overview of the role of human in using an AI product based on the product's impact .....	66
Figure 9: AI Development Pipeline and Sources of Bias (Baker and Hawn, 2021) ..	76
Figure 10 (above): Sample survey question 2: What is the probability that you will keep your permanent address in the same state during the next 5 years? .....	95
Figure 11 (above): Sample survey question 3 .....	96
Figure 12: Distribution of scores for each feature used in training data.....	96
Figure 13a: Gender distribution in the training data of prediction models .....	98
Figure 14a (above): Traders' P & L after rebates .....	109
Figure 15a (above): Auto Correlation plots Contribution Per Lot.....	113
Figure 16a (above): Profit and Loss monthly difference (\$ on y-axis) .....	116
Figure 17: Different paths of a Machine Learning Development Pipeline.....	125
Figure 18a (above) : Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.....	126
Figure 19: Perceived (Blue) vs Actual (Pinks) Agreement between the tool and OMs .....	156
Figure 20 (above): Three Tiers of Transparency .....	167
Figure 21 (above): Three Tiers of Transparency for AI in Education.....	188
Figure 22: Transparency in relation to the other dimensions of ethical AI as depicted in the Transparency Index Framework presented in this research.....	197

## List of Tables

Table 1: Sample survey question .....	95
Table 2: Recruitment tool’s prediction categories and classes for candidates .....	101
Table 3: Exploratory Analysis of traders’ performance .....	111
Table 4: Augmented Dickey-Fuller Test results for performance indicators .....	112
Table 5 (below): Auto Correlation Plot Analysis .....	114
Table 6: Consecutive Month Differences Analysis for Performance Indicators .....	117
Table 7: Levene Test results for performance indicators.....	119
Table 8: Differences and similarities between predicted and actual results .....	129
Table 9a: Contribution Per Lot Classification with Non-Normalized Data.....	132
Table 10a: Profit and Loss Classification with Non Normalized Data .....	133
Table 11a: Performance Bonus Classification with Non Normalized Data .....	135
Table 12a: Hard Stops Classification with Non Normalized Data .....	136
Table 13: Cluster Classification with Non Normalized Data .....	138
Table 14: Recruitment tool’s prediction categories and classes for candidates .....	141
Table 15a: Accuracy Measures of two classes in Profit and Loss predictions.....	145
Table 16: Summary of how the AI tool was used in different offices for decision making.....	152
Table 17: Tool’s Predictions and Office Managers evaluation of candidates .....	154
Table 18: Office managers’ different types of ‘agreements’ with the tool .....	155
Table 19a: Performance Bonus classification for Indian cohort.....	157
Table 20: Alignment of Recruitment Tool’s Models with Indian OMs as perceived by them .....	158
Table 21a: Performance Bonus classification for Polish cohort.....	158
Table 22a: Performance Bonus classification for Ukrainian cohort .....	159
Table 23: Alignment of Recruitment Tool’s Models in Polish Office for Polish Candidates, as perceived by OMs .....	160
Table 24 (below): Alignment of Recruitment Tool’s Models in Polish Office for Ukrainian Candidates, as perceived by OMs .....	160
Table 25 (below): Three Tiers of Transparency .....	168
Table 26: Mapping of different aspects of the Transparency Index Framework from the literature review and ed-tech tool development process .....	171

Table 27: Three groups of people interviewed for this research.....	172
Table 28 (below): Assumptions / Characteristics of the Three Tiers of Transparency .....	189
Table 29: Practical requirements for the Three Tiers of Transparency .....	189
Table 30: Description of the features used and academic references for each feature for training the recruitment tool's models for an ed-tech company in financial services .....	260

# Chapter 1: Background and Research Issue

## 1.1 Introduction

The transparency of AI systems is a rapidly developing field with no agreed upon definitions, or limits. This situation arises because the scope of transparency in AI systems is enrooted in subjective terms like artificial intelligence and ethics which have inherently blurred definitions with no universality. Considering the multi-disciplinary nature of both AI and ethics, these terms can be interpreted in many ways (Weller, 2017; Felzmann, 2019).

Intelligence itself has been defined in at least 70 different ways (Legg and Hutter, 2007), making transparency within the context of AI even more complicated to define. In this thesis, I take a broad definition of AI as being: *any computer system that can interact with the world through capabilities (for example with vision, text and audio) and intelligent behavior (for example processing past information and taking contextualized decisions) that we would consider as requiring intelligence if being completed by humans (Luckin et al, 2016).*

I take ethics to mean:

*a 'rational and systematic study of the standards of what is right and wrong, and morality as the commonly used term for notions of good and bad' (Kazim, 2017).*

Considering the breadth of these definitions of AI and ethics, AI ethics encapsulates principles of philosophy, computer science, engineering, mathematics, politics and economics (Kazim and Koshiyama, 2020; Jobin et al, 2019). It has been divided into various sub-sections like transparency, explainability, fairness, safety, accountability and privacy of AI systems to take account of all the diverse dimensions of ethical AI (Siau and Wang, 2020; Kazim and Koshiyama, 2021).



Developing an AI tool is a complex, time consuming and resource-intensive process. The very first decision to build an AI tool and define its operations involves assumptions that can be challenged or changed. Irrespective of the sector in which AI is applied, transparency is essential to enhance the understanding of relevant stakeholders regarding questions like how the AI works, what are its limitations, in which contexts should it be avoided and how does it improve the status quo.

Mishaps in AI systems are not new. Unintended consequences of AI systems can have a life-changing impact on their stakeholders in certain contexts like education (Tahiru, 2021) where AI has been used for predicting student drop out (Milliron et al, 2014; Christie et al, 2019), graduate level admissions (Waters and Miikkulainen, 2014), knowledge inference (Ritter et al, 2016), essay scoring (Ramineni and Williamon, 2013), tracking collaborative learning (Cukurova et al, 2020; Aldowah et al, 2019; Kent and Cukurova, 2020), visualizing student progress based on pre-determined learning pathways, recommender systems to offer relevant content and adaptive systems to offer personalized content (Long and Siemens, 2011; Wolff et al, 2013; Nistor et al, 2015; Papamitsiou and Economides, 2014).

Considering these several different dimensions within education where AI is having a huge impact, the mishaps of AI within education are not as well documented (Pringle et al, 2016; Paquette et al, 2020) as they are in other sectors like healthcare (Gerke et al, 2020), judicial system (Bennett and Keyes, 2020), recruitment (Pena et al, 2020) or financial services (Zierau et al, 2021). Recently, there has been some work on ethical AI in Education. Berendt *et al* (2020) have discussed the importance of the ethics of education to ensure ethical AI for education. Agudo-Peregrina *et al* (2020) have shown how predictive analytics in AI can be used in virtual learning environments to identify any correlation between different learning constructs and learning outcomes. Elbadrawy *et al* (2016) have personalized learning analytics of students to predict their performance.

Frequently, AI implementations in education assume that the education system is working perfectly and these systems end up strengthening the status quo with all its limitations (West, 2017). Machine Learning (ML) algorithms trained on such data confirm to existing practices and biases (Custer et al, 2018). Adaptive tutoring systems rely on tracking students' progress and actions on their platforms to provide more contextualized learning recommendations. But these tools can grow into aggressive tracking systems which can be used for applications like tracking citizens (Sellgren, 2018, Jack, 2018). AI systems also exert certain amount of influence on learners' choices which can have a significant impact on their life. For example, recommending jobs or courses based on academic performance (Berendt, 2017). The implications of AI systems on education are not thoroughly documented because they are unanticipated and difficult to track except in some case studies (Prinsloo and Slade, 2016; Prinsloo and Slade, 2017).

This research aims to cover this gap of lack of transparency in AI for education by firstly, conceptualizing transparency for AI implementations in educational contexts, and secondly, presenting a framework to facilitate transparent implementations of AI in education.

## **1.2 Ethical AI**

AI is hugely impacting the way we learn (Luckin, 2018), stay healthy (Hansel et al, 2015), cure diseases (Shen et al, 2019), spend money (Smith and Linden, 2017), maintain order in our societies (Brayne and Christin, 2020) and take organizational decisions (Jarrahi, 2018; Philips-Wren, 2012; Algorithm Watch, 2019). This penetration of AI in our daily lives has also magnified the risks it poses (Andrew et al, 2019).

There is a growing focus on AI ethics to raise awareness on the pitfalls of AI with its lightning spread across various dimensions of our lives. At least 84 different public-private initiatives have produced various principals, tools and design frameworks to guide the ethical development of AI (Mittelstadt, 2019; Greene et al, 2019). Considering the complexity of the AI development

process, ethical AI has been divided into different dimensions to address the different needs of trustworthy AI (Floridi, 2019). These dimensions include transparency, fairness, explainability, accountability, privacy, safety and interpretability. There is an overlap between these dimensions and at times the boundaries between them seem blurred. On top of this, there is no consensus on the definitions of these dimensions or ethical AI in general. Different tools and frameworks address different dimensions of ethical AI (Brundage et al, 2020; Dameski, 2018; Hagendorff, 2020; Siau and Wang 2020).

Partnership on AI, an organization that brings together AI researchers and practitioners from around the globe found “a gap between explainability in practice and the goal of transparency, since current explanations primarily serve internal audiences, rather than external ones”<sup>1</sup> (Bhatt et al, 2019). John Danaher (2018) has shown that the ethics of AI-powered personal assistants (like in other contexts) is complex and not dependent on one particular decision, tool or framework. This can also be applied to education where a number of Intelligent Tutoring Systems (Kim et al, 2020) and AI-powered learning assistants have emerged. But they have pedagogical limitations (Herold, 2017; Watters, 2015; Watters, 2017; Wilson and Scott, 2017). Alonso and Casalino (2019) have shown the effectiveness of AI explanations in Virtual Learning Environments (VLEs). But they also show that these explanations need to be understandable for the learners and to make it understandable, interpretability and transparency can play an important role.

Each of the dimensions of ethical AI have their risks associated with AI’s applications in education. Management, usage and storage of learner data is covered by privacy, lack of bias or discrimination is covered by fairness, reasoning and logic behind AI system’s results is covered by explainability and interpretability, penalty for mistakes by AI systems is covered by accountability and getting a deeper understanding of how an AI system is built and performs is covered by transparency (Chaudhry et al, 2022). Putnam and

---

<sup>1</sup> <https://www.partnershiponai.org/xai-in-practice/>

Conati (2019) have shown how two different types of explanations can be used in Intelligent Tutoring Systems (ITS) to facilitate learners. The effectiveness of these explanations is dependent on how well they are perceived by the learners.

All the dimensions of ethical AI mentioned above are extremely important for AI's applications in education. There is a lot of overlap between them, but each of them addresses different aspects of the unintended consequences of AI. For example, transparency and interpretability can help in increasing the understandability of clustering algorithms when applied to education.

Clustering is one of the AI techniques or ML algorithms that have been extensively used in AI's applications to education (Dutt et al, 2017; Vandamme et al 2007; Bresfelen et al 2008). It is considered one of the most important unsupervised techniques to sort out data between similar and dissimilar features (Madhulata, 2012). In education this AI technique has been widely used to group together learners with similar learning needs or pathways (Wise et al, 2013; Ivan cevic et al, 2012; Zajac et al, 2019, Chen et al, 2007). But extensive research in learning sciences show that each learner is unique and may have different learning needs (Prain et al, 2013; Maseleno et al, 2018). This does not imply that clustering techniques should not be applied in educational contexts. In fact, the divergence between learning sciences and clustering algorithms has been addressed by making their implementations transparent and easily understandable in the context of education (Purba et al, 2018). It involves highlighting the decisions and assumptions made in grouping the students into different clusters and critically evaluating when these clusters can mislead educators (Li et al, 2020).

This research focuses on transparency as a necessary construct of ethical machine learning powered AI that can enhance human understanding of complex AI systems and enable the implementations of other dimensions of ethical AI in the AI development pipeline. By machine learning powered AI, I refer to AI products that are powered by machine learning algorithms like

linear regression, logistic regressions, k-means clustering, neural nets or decision trees etc (Ray, S., 2019; Ayodele, 2010). This does not include the rule-based AI systems that have human-made rules to store, sort and manipulate data and produce pre-defined outcomes<sup>2</sup> (Gruntiz, M., 2021; Hayes-Roth, F., 1985).

Researchers at Fujitsu Laboratories like Chander et al (2018) have discussed transparency in AI in the context of 'human in the loop', where the relevant teams have a better understanding of how their AI system works. They discuss the accessibility, explainability, interactivity and tunability of AI under the umbrella of transparency. Fig 1 illustrates the gap between collected data from human experiences and AI's predictions from human beliefs. It shows that human experiences are high dimensional, which means they consist of several factors such as time, context, other humans' involvement and their own perceptions of a particular moment. Beliefs are formed from experiences, but unlike experiences, they are not high-dimensional. This means, in terms of data, beliefs are not dependent on the richness of data, like experiences. The gap between human beliefs and AI's predictions is called the Persuasion Gap and it can be considered as one of the indicators to measure the performance of the model. The importance and validity of Digital Data shown in Fig 1 is variable and dependent on the richness of data collected. The gap between the data and human experiences is called The Awareness Gap.

In education, awareness and persuasion gaps can have a major impact on the usefulness of AI products. For example, for an awareness gap, collected data can usually take account of students past academic performance, their attendance in classrooms or activities on an online learning portal, but it may not take account of the impact of change in their family's conditions due to parents' unemployment, a sibling's illness or their passion for sports on their education. This creates a gap between the digital data collected for an AI powered ed-tech and learner/educator actual experiences. Transparency of

---

<sup>2</sup> <https://wearebrain.com/blog/ai-data-science/rule-based-ai-vs-machine-learning-whats-the-difference/>

machine learning powered AI systems can help in the identification and mitigation of awareness gap.

A persuasion gap occurs when an AI tool predicts a low grade for learner based on their past performance, but teacher believes that the student will perform much better because they have started working hard, are more focused and are taking dedicated personal help from a private tutor. These factors are not taken into account when collecting the digital data, which creates an Awareness Gap in the AI development process. The awareness gap translates into lower accuracy or performance of an AI system, in comparison to human judgment, which creates a persuasion gap.

In some unique cases, under certain simulated environments like the game of Go (Silver et al, 2017) or Starcraft (Vinyals, 2019), data collected from years of self-play in deep reinforcement learning algorithms (Sutton and Barto, 1998; Sutton and Barto, 2018) can produce surprising results for humans (Iyer et al, 2018). These algorithms sometimes take actions that surprise humans<sup>3</sup>.

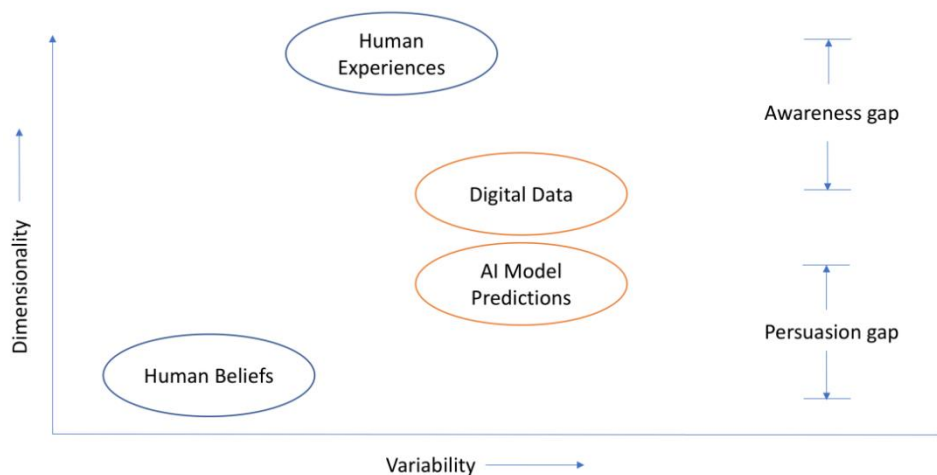


Figure 1: Search for Common Cognitive Ground

In the research conducted as a part of this thesis, the proposed Transparency Index Framework can enable various stakeholders of an AI system to better understand its strengths and weaknesses. This framework can increase the chances of exposing hidden beliefs in the data. It utilizes the Awareness Gap

<sup>3</sup> <https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/>

and the Persuasion Gap to identify, document and report deficiencies across the different stages of AI development pipeline in an easily understandable manner.

### **1.3 Bias**

Bias in AI systems is not a new phenomenon. It is a necessary drawback of AI systems trained on data (collected from society that has bias inherent in its operations) and developed by AI practitioners (who have their own beliefs about the world that can be biased). It directly relates to the lack of diversity and inclusion in AI systems. Ethical AI and its various dimensions such as fairness, accountability or explainability aim to address and mitigate the impact of bias in AI system. Mitchell *et al* (2021) have defined bias in the context of AI as when ‘a model’s predictive performance unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class”. They have divided bias into two broad categories, statistical and societal bias. Statistical bias can occur due to a sampling or measurement imbalance in the data and societal bias occurs due to ‘objectionable social structures that are represented in the data’ (Mitchel et al, 2021) and cannot be counteracted by increasing the quantity of data.

Bias has been commonly found in AI-powered ed-tech products. (Bridgeman et al, 2009; Bridgeman et al, 2012; Ocumpaugh et al., 2014; Yudelso et al., 2014; Kai et al., 2017; Hu & Rangwala, 2020; Yu et al., 2020). Ocumpaugh et al (2014) showed that students’ emotion detectors trained on urban, rural and sub-urban student populations perform better when they are trained on a single group’s data rather than all three populations. Number of researchers have shown that bias exists in using AI for predicting learners at risk of failing a course (Hu & Rangwala, 2020; Lee & Kizilcec, 2020).

In education, the impact of bias in AI can have severe consequences for learners and educators. The differences between AI’s predictions and teachers’ perceptions about a learner can potentially lead to confusion and lack of confidence for teachers. This can cause automation bias as teachers may trust AI’s predictions over their own better judgment (Skita et al, 2020;

Goddard et al, 2020). For learners, an AI wrongly predicting the drop out for a particular student (Anderson et al, 2019) can tarnish their reputation in front of teachers and parents. It would affect how educators treat that learner and can lead to misleading learning pathways and pedagogical choices. The psychological impact of this single prediction can have significant psychological implications, both short-term and long-term for that learner.

Suresh and Guttag (2019) divided the AI tool development process into a 'Data Generation' and 'Model Building and Implementation' stage (shown in figure 2) to identify the different types of bias that can exist in the AI tool development pipeline.

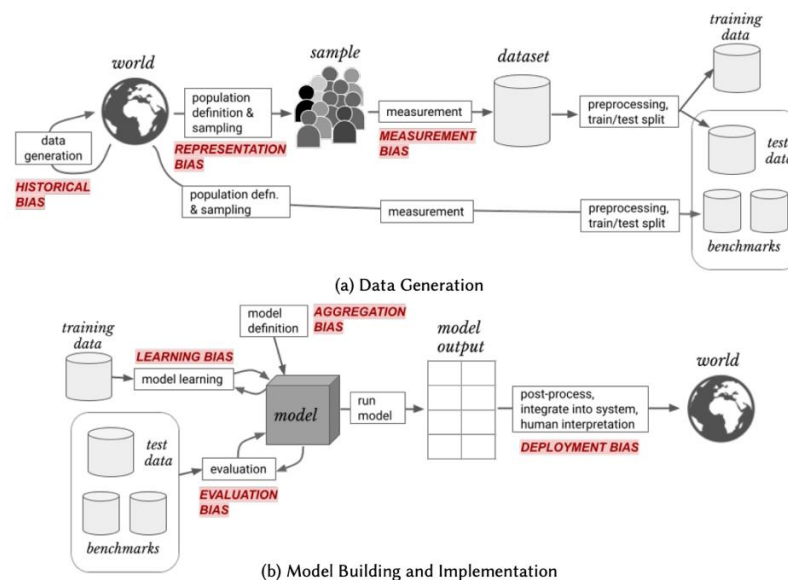


Figure 2: Steps involved in the development of an AI product with different types of biases that can occur in different stages

They reviewed the AI development pipeline strictly from an ethical AI perspective with a focus on different types of biases that can affect the results of an AI tool. Some of the biases that Suresh and Guttag (2019) identified have been presented below:

- **Historical Bias:** occurs in the data collection stage and is enrooted in the real world as it is. It is not dependent on the data processing and model building stages. These are facts that exist in the data and real world. Increasing sample size would not change them. For example, predicting student



admissions for stem courses based on (correctly sampled) data where less than ten percent applicants in the past five years have been females would be an example of a historical bias. The chances of an AI tool recommending a female student would be comparatively lower.

- **Representation Bias:** occurs in the data processing stage when the selected sample is not representative of the real population. Data might be skewed towards a certain group or away from a certain group. For example, if the training data for university admissions consists of eighty percent white, male applicants from north of England between seventeen and nineteen years old, the chances of an AI tool recommending someone outside this group would be much lower.
- **Measurement Bias:** occurs in the data processing stage when choosing or shortlisting the features of interest from the sample population. There can be bias in the measurement of these features of interest. For example, the A-level grading fiasco in UK (Kippin and Cairney, 2021; Jackson and Panteli, 2021) took account of historical grade distribution of past three years from schools. This meant that from a particular school, if students have not performed well in the past three years, they are unlikely to perform well this year too<sup>4</sup>. This reflects a measurement error when collecting data for training an AI tool.
- **Evaluation Bias:** occurs in the model building stage during a model's evaluation and iteration when model's parameters are biased. For example, an AI powered ed-tech tool trained on data from English, white, male students at primary level is applied on Asian, Chinese, female students or if the performance of this tool on an underrepresented group in the data is kept a secret. This bias can occur due to any of the historical, representative or measurement biases mentioned above.
- **Aggregation Bias:** arises due to wrong assumptions about the population and can persist throughout the data processing and model building stages. For example, a natural language processing tool analysing students'

---

<sup>4</sup> <https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/>

communication on online portal might categorise students based on its pre-conceived notion of emotions like helpful, rude, angry, frustrated etc. In such contexts, students with English as second language may be categorised incorrectly due to cultural differences,

The goal of an AI system in real-world is usually to produce as accurate results as possible, with none of the biases mentioned above. Some of these biases may overlap, and multiple biases can exist in a machine learning setting. Addressing these biases requires a strong commitment and considerable amount of time and resources from the companies and leaders spearheading the development of AI tools. Firstly, to diagnose, and secondly, to mitigate the effects of these biases. Hence, they form an important part of the Transparency Index Framework presented in this research.

All the biases above require different diagnostic and mitigating techniques (Mehrabi et al, 2021). Identifying and tackling each of these biases separately can be a tedious and time-consuming process for the AI practitioners and project managers who are mostly working on tight deadlines. One solution does not fit all in this case, and a particular tool cannot be used to address all these biases. Solutions are also dependant on the context in which the tool is being developed, and in which it will be used. For example, what if a researcher builds an AI system that does laughter identification through facial recognition, and it is biased against women. Let's suppose it gives a lot of false negatives when dealing with women's faces. Researchers go back to the data collection process, get more samples of women's faces and retrain the model. This can remove the bias from their system. This was a Representation Bias because the researcher's data lacked women's representation.

On the other hand, imagine an AI-powered recruitment tool that predicts the suitability of a candidate for a particular role. The AI practitioner has a dataset of the candidates with human assigned ratings of who would be suitable for the role and vice versa. The developer notices that the algorithm's likelihood of recommending women for that role is much lower so the data processing

stage is revisited and a lot more data from female candidates is collected. This still does not solve the problem, and bias persists. Contrary to the earlier case, the problem in this scenario is not insufficient data from a particular group, hence collecting more data from female candidates from the distribution does not help. The issue, in this case, is the human label that has been used to determine the suitability of a candidate. This can be categorized as a measurement bias when certain assumptions about the sample population while doing feature selection are taken.

The above example highlights the complexities involved in the AI tool development process and the importance of transparency in not only understanding the problems faced in designing and developing AI tools, but also making sure that the development process is robust, fair and inclusive. Bias in AI systems is a cause as well as an effect of the lack of inclusion and diversity in AI systems. Bias can be caused by lack of diversity and inclusion of sensitive features like gender, race, ethnicity, geographic location etc in the training data or by the lack of diversity of such features among the team that is designing and developing an AI-powered ed-tech.

The solution to tackling bias in AI systems depends on the type of bias, where it occurs in a machine learning pipeline, what are the real causes behind it and the priorities of business leaders spearheading the development of an AI tool. For example, they may want more true positives compared to false negatives in results. These factors also affect the time it would take to identify a bias and then addressing it. For example, historical bias inherent in the data may take more time to be identified than a representation bias. It also depends on the metrics that are achieved by tackling that bias and what business leaders prefer, or researchers and AI practitioners recommend. For example, if equal metrics for different groups (Corbett-Davies et al, 2017) through true positives (Hardt et al. 2016) are preferred or if a reduction in false negatives is the goal. These performance goals for an AI system made during the development of an AI tool can play a very important role in determining the tool's performance in different contexts. These choices can augment the biases in an AI system or mitigate them. Hence their documentation, tackling

strategies and implementation tools should be a part of any initiative taken to make the AI development process more transparent, especially in educational contexts where bias can have a huge impact on stakeholders (Baker and Hawn, 2021).

At times the data that goes into an AI system has inherent societal bias (Caliskan and Narayanan, 2017) irrespective of the sample size. Tackling such bias goes 'beyond technical debiasing' techniques (West et al, 2019). No ethical AI tool or framework can guarantee a 'perfectly fair' or 'perfectly explainable' AI system for everyone because definitions of fairness may change (Chouldechova, 2017) and explainability of an AI system is subjective and dependant on the targeted users' understanding (Gade et al, 2019). This means the choices that AI practitioners make in mitigating AI tools' biases, improving tools' accuracy metrics or making the tools explainable will only work under certain conditions, in certain contexts and for selected number of stakeholders (Durmus, 2022). This highlights the importance of transparency for documenting and sharing all the decisions and assumptions made during the AI tool development process.

#### **1.4 Why Transparency?**

Recently, there has been a lot of research and adoption of ethical AI principles like fairness, accountability, interpretability and explainability (European Commission, 2020; IEEE, 2019). This is driven by the impact of AI applications in our daily lives (Bughin et al, 2018; Crawford et al, 2016; Vaishya, 2020), and the mishaps of AI systems in the real world (Kaushal et al, 2020; Wellner, 2020; Raji and Buolamwini, 2019; Završnik, 2020). There have been a number of tools, checklists and frameworks published to take account of ethical considerations in developing, deploying and auditing AI products (Morley, 2020; Dameski, 2018; Leikas, 2019; Winfield, 2019; Deepmind Safety Research, 2018) but there is not much work focusing on transparency in particular throughout the machine learning powered AI's planning, development and deployment stages.

An AI tool built with huge amounts of data and the best performing machine learning algorithms (Fedus et al, 2021; Brown et al, 2020) will perform at its best only in certain contexts. Transparency is essential to know in which contexts the tool will not perform at its optimal level. It is widely accepted that bias or discrimination cannot be completely removed from an AI tool (IBM, 2018; ICDPPD, 2018), but we can mitigate them. The extent to which bias or discrimination exists in a particular AI product, applied in a certain context with a particular type of users can only be determined if the details of the tool's development are documented and shared.

Richard and King (2013) have identified transparency among the three paradoxes of big data in AI. AI aims to make the world more transparent and AI tools in different sectors like healthcare, education and governance claim to empower individuals, but they seem to be doing this in a very secretive manner where decisions and assumptions are not documented (Holstein et al, 2019), machine learning models used are opaque (Castelvecchi, 2016) and the limitations of these tools are not shared (Besold, 2014).

Ananny and Crawford (2016) have discussed the limitations of transparency in ensuring or guaranteeing ethical AI. They evaluate transparency at two levels: algorithmic transparency in the code that AI practitioners write (Pasquale, 2015; Diakopoulos, 2016; Brill 2015; Dubber et al, 2020) and design transparency regarding how the AI systems are planned, developed, deployed and evaluated (Hollanek, 2020, Plale, 2019; Wischmeyer, 2020). Some of the issues raised by these researchers are discussed after presenting the Transparency Index Framework later in this thesis.

There are three main reasons why there has been an increase in the demand for transparent machine learning powered AI products and tools: Firstly, the impact of AI on society and the day-to-day living has increased dramatically in the past few years (Makridakis, 2017). Some of the tools used daily like Amazon to shop online, Uber to commute, Google to search for information or online platforms to find relevant learning content are powered by AI to

optimize experiences. These AI-powered tools have become an integral part of daily lives.

Secondly, even though AI has a huge impact on daily lives, sometimes even researchers and practitioners don't know how the black box works (Castelvecchi, 2016; Tan et al, 2019) or how to make it explainable (Adadi and Berrada, 2018). A lot of research is being directed towards the opening of this black box to enable users to know why an AI is taking certain decisions. Some researchers like Yang and Kandogan (2019) have proposed a 'human in the loop' approach where a human is actively involved in AI decision making to make this process more transparent and explainable.

Thirdly, there have been an increasing number of cases where AI development and deployment has gone wrong and has adversely affected its users (Yampolskiy and Spellchecker, 2016). In the last few years an increasing number of Artificial Intelligence failures<sup>5</sup> all around the world in sectors such as healthcare (Obermeyer et al, 2019), recruitment (Dastin, 2018), justice system (Angwin et al, 2016), chatbots (Perez, 2016) and education (Kippin and Cairney, 2021) have brought transparency at the centre of AI development and deployment process. These were enterprise level failures that had a huge impact on the lives of number of people. There have also been other mishaps in AI systems that have raised questions on their reliability, like an Uber self-driving car killing a pedestrian (The Economist, 2018), personal assistant misunderstanding a child's voice command and playing porn<sup>6</sup>, a facial recognition being tricked by a plastic mask<sup>7</sup> and an algorithm predicting wrong grades of A level students<sup>8</sup> (Satariano, 2020). Such incidents have made transparency and ethics a focal point for discussion in AI development. There is no doubt that the potential of such algorithms to cause harm would increase as AI is adopted more widely

---

<sup>5</sup> <https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2019>

<sup>6</sup> <https://www.entrepreneur.com/video/287281>

<sup>7</sup> <https://www.wired.com/story/hackers-say-broke-face-id-security/>

<sup>8</sup> <https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html>

(Bostrom, 2014) and as the perception of Machine Learning algorithms improve with time (Russell, 2015).

## 1.5 Research Questions:

The overarching research question at the heart of this thesis is:

- *What design framework can be applied to ensure an optimal level of transparency in machine learning powered Artificial Intelligence (AI) products in educational contexts?*
  - *How should an optimal level of transparency be conceptualised for different stakeholders of machine learning powered AI in Education (AIED)?*
  - *How can existing frameworks for ethical AI be applied in the context of education for transparent AI development pipelines?*
  - *How can a design framework assist in addressing and understanding the Awareness Gap (the gap between digital data and human experiences) in AI-powered ed-tech?*
  - *How can a design framework be utilised by different stakeholders to make more informed decisions regarding the selection and development of AI-powered ed-tech?*

This research is at the intersection of Learning Sciences (LS) community and Machine Learning (ML) community. Researchers from both communities have raised the issue of lack of understanding of each other's work<sup>9</sup> and a difference in alignment of their goals (Fiok et al, 2021). ML researchers and AI practitioners working in educational contexts need to know the nature of learning and how learners learn, while learning scientists need to have a much better understanding of the data used for ML modelling (Jacobson et al, 2019) and other AI development details that may impact its decisions.

This PhD brings Artificial Intelligence and Education (AIED) together by bridging the gap between ML and LS researchers through its proposed

---

<sup>9</sup> <https://www.tonybates.ca/2019/11/08/learning-analytics-in-online-learning-trying-hard-but-need-to-do-better/>

framework. For ML researchers and practitioners, the Transparency Index Framework (TIF) proposed in this research offers a toolkit and a checklist to make their AI-powered products more accessible and understandable for the LS community. For LS researchers and educators, TIF offers a checklist to audit the AI-powered products and provides guidelines on what kind of questions Learning Scientists need to ask from AI-powered ed-tech providers or what information they need to enhance their understanding of AI-powered products. Figure 3a highlights the research questions this research address, the gaps in the literature it fills and how it contributes to and advances the field of AI in education. The final version of the Transparency Index Framework presented in this research was developed in three stages using a mixed methods approach: firstly a thorough literature review of different AI development pipelines and ethics frameworks was conducted, secondly the shortlisted frameworks were applied in the AI development process of an AI-powered ed-tech tool, and thirdly, the Transparency Index Framework was iteratively improved based on the direct feedback from different stakeholders of AI in education.



	Main Research Question: What design framework can be applied to ensure optimal level of transparency in machine learning powered Artificial Intelligence (AI) products in educational contexts?			
Research Questions	RQ 1: How should an optimal level of transparency be conceptualised for different stakeholders of machine learning powered AI in Education (AIED)	RQ 2: How can existing frameworks for ethical AI be applied in the context of education for transparent AI development pipelines	RQ 3: How can a design framework assist in minimising and understanding the Awareness Gap in AI powered ed-tech?	RQ 4: How can a design framework be utilised by different stakeholders to make more informed decisions regarding the selection and development of AI powered ed-tech?
Gaps	Gap: Conceptualization of transparency for AI in Education: theoretically and pragmatically for stakeholders like educators, ed-tech experts and AI practitioners	Gap: A theoretical and empirically based framework for transparent machine learning powered AI development pipelines in educational contexts	Gap: Exploration of factors that address the Awareness Gap in educational contexts for AI products	Gap: Evaluation of a design framework for transparency in AI powered ed-tech products through different stakeholders of AI in Education
Goals	Goal: Contextualize transparency for the diverse set of stakeholders of AI in Education including ed-tech users and companies	Goal: Build a framework for transparency in machine learning powered AI products in educational settings, based on existing ethical AI frameworks	Goal: Empirically and theoretically detect the factors from users (learners, educators and ed-tech experts) and suppliers (ed-tech companies) perspective that impact the awareness gap	Goal: Evaluate and improve the framework based on stakeholders' feedback including educators, ed-tech experts and AI practitioners working on ed-tech products
Methodology	Methodology: Thorough literature review of Transparency in AI and Transparency in AI for AIED and interviews with different stakeholders of AI in Education	Methodology: Literature review to identify different ethical AI frameworks and then real-world implementation of these frameworks in educational contexts	Methodology: Practical implementation of ethical AI frameworks in an educational context and interviews with different stakeholders of AI in Education	Methodology: Interviews with educators, ed-tech experts and AI practitioners to incorporate their feedback in Transparency Index framework
Contributions	Contribution: An interpretation of Transparency in accordance with the different tiers of stakeholders of AI in Education	Contribution: Application and alteration of exiting ethical AI frameworks (for different stages of the AI development process) to suit educational contexts	Contribution: A Transparency Index Framework for machine learning powered AI in Education covering the entire AI development pipeline	Contribution: Insights into the requirements of educators, ed-tech experts and AI practitioners for transparent AI powered ed-tech products

Figure 3a: Research question and contributions of this research

## 1.6 Methods

Considering the inherent complexity of most educational contexts, evaluating ed-tech tools or identifying the need for frameworks or guidelines to inform the ed-tech tools usually requires multiple or mixed methods (Mark and Shotland, 1987; Mertens, 2005; Scott and Sutton, 2009; Ponce and Pogan, 2015).

Mixed methods approach was used to develop, evaluate, and improve the framework proposed in this research as shown in figure 3.

Firstly, the initial version of the framework was developed in two phases based on a thorough literature review of the standard machine learning development pipelines for AI systems (phase 1) and an application of shortlisted frameworks on different aspects of the AI development process for educational contexts (phase 2). In phase 3, data driven approach was used to evaluate and iteratively improve the framework, based on the qualitative data collected from the interviews of different stakeholders of education. Figure 3b shows the iterative cycles of three phases including the tasks undertaken in each phase and outcomes achieved.

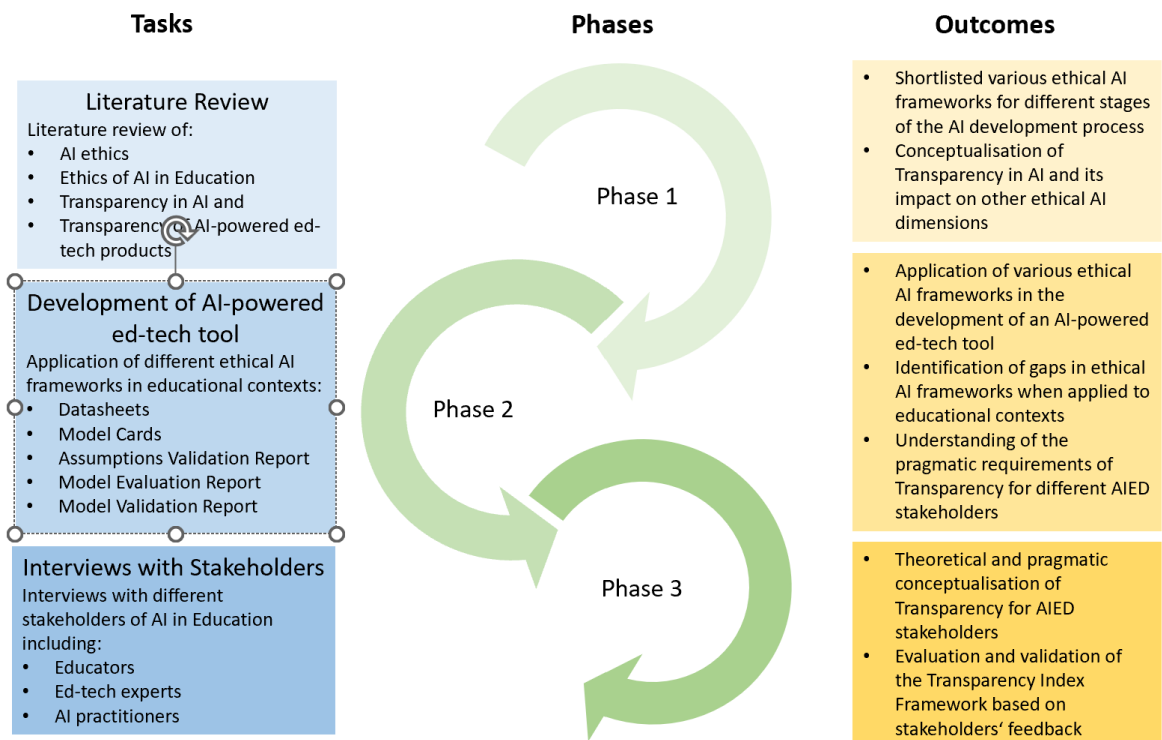


Fig 3b: Three iterative phases of the participatory co-design process used for developing, evaluating and validating the Transparency Index Framework

### **1.6.1 Framework Creation Methodology**

The framework was created and co-designed in three phases: firstly, based on a literature review of different stages of the AI development process. Secondly, the application of selected frameworks (from phase 1) on the development of an AI-powered ed-tech tool for a financial services company. Thirdly, the interviews with different stakeholders of AI in education to iteratively explore its relevance and value to stakeholders and improve the framework accordingly.

#### **1.6.1.1 Ethical Considerations**

Ethics approval for this whole study was granted by the Institute of Education, University College London. The ethics approval form has been added in Appendix 7. Informed consent was taken in two different phases of the research. Firstly, from the employees of the financial services company whose data was used in phase two of this study. This informed consent was taken by the financial services company before sharing the data with my research team. Secondly, from all the participants who were interviewed in phase 3. In phase 2, all the traders' data was pseudonymized and then shared with the research team. The research team consisted of a project lead, an AI practitioner (myself), a learning scientist and a research lead. The conversational and other data from the office managers were collected by the project lead and analysed by me in collaboration with the research team. The data used for the model validation report in chapter 5 was pseudonymized and shared by office managers in Excel sheets. This data was analysed by me for validating the AI-powered ed-tech tool. The informed consent form used in phase 2 is given in appendix 4, the informed consent form used in phase 3 is given in appendix 5 and the interview questions for phase 3 (stage 1 and stage 2) are given in appendix 6. Further details of ethical considerations have also been discussed in the methodology section for each

phase: section 1.6.1.2 (for phase 1), section 1.6.1.3 (for phase 2) and sections 1.6.2 and 6.2 (for phase 3).

### 1.6.1.2 Phase 1

In phase 1, the literature on the development processes of AI-powered products, ethical considerations and frameworks for the pipeline development of AI-powered products, the use of AI-powered products for educational contexts and ethical AI in education was reviewed.

In this first phase, the chronological order of the three different stages involved in the development of an AI-powered product were identified. Then, for each stage of the development of an AI tool, ethical frameworks for documentation, robustness and reproducibility were identified. The first phase formed the basis for the methodology adopted for developing an AI-powered ed-tech tool in phase 2. During phase 1, the meaning of transparency and how it may inform the other dimensions of ethical AI like explainability, fairness, accountability and safety were also explored.

For the data processing stage, datasheets from Gebru et al (2018) were chosen as the benchmark for documenting different components of the data processing stage in the AI development process. In the Transparency Index Framework, datasheets were wrapped around other requirements for the data processing stage to make it more applicable and suitable for educational contexts.

For the Machine Learning modelling stage, model cards by Mitchell et al (2019) were chosen as a baseline to document the details of the ML model used. Some additional requirements were also added for this stage to record the various decisions and assumptions made specifically for the AI-powered ed-tech products. These requirements were derived from my experience of applying AI in educational contexts to enhance the understanding of domain experts of the trading behaviour and performance of the traders they recruited (Kent et al, 2021).

Factsheets by Arnold et al (2019) were added as a basic requirement for documenting the usability of AI tools. Considering the importance of user feedback in ensuring the effectiveness of AI tools in educational contexts, some additional requirements were also merged with factsheets for the deployment and testing of AI-powered ed-tech products in real-world settings. For example, preparing a model validation report for the AI-powered ed-tech tool and documenting any endorsements attained from regulatory bodies.

AI development is a complex and time-consuming process. All these frameworks cover different stages of the AI development process. They were brought together in a coherent manner under the umbrella of the Transparency Index Framework built as part of this research study and were accompanied by some other requirements to specifically suit the needs of different stakeholders in education.

#### 1.6.1.3 Phase 2

In the second phase, the selected frameworks for each stage of the AI development process were applied in the real world during the development of an AI tool in educational contexts for a financial services company that envisioned to become a leader in educational technology for financial services. This application of domain-agnostic frameworks for AI tool development in any context or domain enabled me to identify the gaps in these frameworks, when applied in educational settings.

The AI-powered ed-tech tool prepared for a financial services company during this research was designed with a 'human in the loop' approach. The company specializes in electronic futures trading in different asset classes and looks for candidates who can succeed in these trading jobs. The purpose of this project was to utilize AI to augment human intelligence. The role of the ed-tech tool was to educate the office managers and recruitment managers to take more informed decisions when selecting potential candidates for trading jobs. The tool provided an extra set of insights powered by AI to decision

makers (who were office managers), based on all the data of current and former employees.

This was an ed-tech tool aiming to enable more informed decision making by providing predictions about prospective candidates trading performance based on the company's former employees. The AI tool did not provide any conclusive recommendations to learners (office managers) or made any decisions on behalf of the office managers, for example, whether to recruit or not to recruit a candidate. Rather, the tool visualized the predictions for certain trading performance metrics like the potential these candidates have for driving profit after nine months, their contribution to the company and the performance bonus they can potentially earn. Learners (office managers) used this information along with their own judgment of candidates' potential to make the final decision of hiring or not hiring a candidate.

In phase 2, for the data collection and processing stage of the AI development process, I prepared the Datasheet (shown in chapter 3) for the AI-powered ed-tech tool. Along with highlighting the various aspects of the data collection and processing stage, it also documented the ethical measures taken during the data collection process (phase 2) of this study, including the informed consent that was taken from all the employees of the company whose data was used in this research for knowledge elicitation, training the ML models and validating the ed-tech tool. After collection, the data was pseudonymised so that no individual could be identified. All the data collection, storage and processing were done according to the GDPR regulations. In the data collection and processing stage, office managers highlighted some of the assumptions based on which the recommended data was collected. The process of documenting and validating the assumptions (shown in section 3.3) was added as an additional requirement to the Transparency Index Framework.

For the ML modelling stage during phase 2 of this study, I prepared a Model Card (shown in chapter 4) for the machine learning model used to make the predictions about traders' expected performance. During the integration of

transparency considerations into the ML modelling stage of the AI development process, I noted that Model Cards documented the details of the model but did not take into account the process I followed in selecting a particular machine learning model for the AI-powered ed-tech tool. Hence, the Models Evaluation Report (shown in section 4.2) was added as a requirement in the Transparency Index Framework.

For the AI deployment stage during phase 2 of this research, I prepared the model validation report (shown in chapter 5) to test and evaluate the value and usefulness of the AI-powered tool. During this stage of the AI tool development process, data was collected (by the project lead) through interviews and an Excel sheet (illustrating the number of times office managers considered the tool to agree with them) from four office managers as the feedback on how they used the tool and which aspects of the tool were helpful or confusing. Two office managers used the tool in Poland and Ukraine and the other two used the tool in India. The model validation report focused on the details of how the AI-powered ed-tech tool was used but did not take into account the trust and transparency of the tool. To take these into account, I suggested Factsheet as a requirement for transparency in the AI deployment stage during the development of AI-powered ed-tech.

The project for developing an AI-powered ed-tech tool during phase 2 of this research was a joint project with the financial services company. I noted that to make the AI tool development process transparent for different stakeholders in the financial services company, different measures are required. I maintained a diary of these groups of stakeholders with the measures taken to make the AI-powered ed-tech tool transparent for them.

As detailed above, in every step of the development pipeline further improvements were suggested and added as contributions of this research study to the initial approaches identified from the literature reviewed.

Therefore, the first version of the Transparency Index Framework was prepared based on phase 1 (literature review findings on relevant

approaches) as well as phase 2 (AI-powered ed-tech tool development experience) specific to this research project.

### **1.6.2 Framework Evaluation Methodology**

In phase 3, three different groups of stakeholders of AI in education were consulted in two stages through interviews for evaluating the usefulness of the framework. The stakeholders consisted of educators, ed-tech experts, and AI practitioners. In stage one of phase 3, ten stakeholders were interviewed, and the first version of the framework was shared with them for feedback. The framework was revised based on stage 1 feedback. In stage 2 of phase 3, eight more stakeholders were interviewed to confirm the value and effectiveness of the revised framework and make further improvements accordingly.

For phase 1, chapter 2 highlights the context of this research study and presents the literature on the AI development process, conceptualisation of transparency, ethical AI frameworks, the use cases of AI in education, and the ethical use of AI in education. For phase 2, chapters 3, 4 and 5 follow the chronological order of an AI development pipeline with chapter 3 presenting the transparency considerations for the data processing stage, chapter 4 presenting the transparency considerations for the machine learning modelling stage and chapter 5 presenting the transparency considerations for the AI deployment stage. For phase 3, chapter 6 presents the iterative co-design of the final version of the framework through interviews with educators, ed-tech experts and AI practitioners. Chapter 7 presents the final version of the framework along with a discussion on the concept of transparency in the context of AI in education. Lastly, chapter 8 highlights the contributions of this research along with its limitations and potential future work.

## **1.7 My role in the project**

In phase 2 of this research study, I was part of a bigger research team who conceptualized and prototyped the AI-powered ed-tech tool. Five different



reports from that project have been presented in this thesis as part of the phase 2 methodology, comprising chapters 3, 4 and 5. My specific role and contributions to each report are presented below:

1. Datasheet: I prepared the datasheet presented in section 3.2 of this thesis. In the datasheet, I documented various aspects of the data collection, storage, analysis and compliance with regulations like GDPR.
2. Nine-Month Analysis Report: I prepared the Nine Month Analysis Report presented in section 3.3 of this thesis. The purpose of this report was to test the assumption that traders' performance stabilizes after nine months of joining the financial services company. This report was improved based on the feedback from the other members of my research team.
3. Models Evaluation Report: I prepared the Models Evaluation Report for this project presented in section 4.2 of this thesis. The purpose of this report was to experiment with different ML models on the collected data. This report played an integral role in choosing the ML model used for developing the AI-powered ed-tech tool.
4. Model Cards: I prepared the Model Card for the ML model used in the ed-tech tool. This is presented in section 4.3 of this thesis. The final training and evaluation of the selected model were done by the project lead of the team. I documented the whole process in the Model Card.
5. Model Validation Report: I prepared the Model Validation Report for the ed-tech tool that has been presented in section 5.2 of this thesis. This report was improved based on the feedback from other members of the research team. For this report, interview data from office managers was collected by the project lead. I collected and analyzed the Excel sheet data which highlighted how office managers perceived their learnings and predictions from the tool.

## **1.8 Conclusion**

Transparency as conceptualised in this research can play a pivotal role in ethical AI development for educational contexts. The different chapters of this thesis present different phases of the Transparency Index Framework's development process. Chapter 2 highlights the literature review conducted in phase 1. Chapters 3, 4 and 5 present phase 2 during which an AI-powered ed-tech tool was developed and chapters 6 and 7 highlight the evaluation and co-design of the framework with various AI in education stakeholders. Lastly, chapter 8 presents the limitations of the Transparency Index Framework and possibilities of future research.

## **1.9 Glossary**

Following are the key terms and their definitions used in this research study:

1. Activity of the trader on the trading system: number of trades submitted by a trader on the futures trading platform.
2. AI-powered products: educational technology tools that have machine learning-powered AI built into them.
3. AI tool development process: different stages involved in building an AI-powered product, such as data processing, ML modelling and deployment.
4. Augmented Dickey-Fuller Tests: a statistical test used to test if a given time series is stationary or not.
5. Auto-correlation: is the similarity between different observations as a function of the time lag between them.
6. Bias: prejudice against a group considered to be unfair.
7. Child order: a new trade under an already existing trade that was executed in the past.
8. Clusters: grouping of traders based on their trading behaviors taking account of data from the following factors: activity of the trader, volume and held positions, diverse vs complex trading and volatility vs liquid market preferences.
9. Contribution to the company: money earned by the financial services company based on the trades executed by the trader.

10. Diverse vs complex trading: trading in multiple (many) products as compared to specializing in trading a few products.
11. Few shot learning: a technique/framework that enables a pre-trained ML model to generalize to new tasks with a few labelled samples.
12. Hard stop counts: the number of times a trader reaches the daily trading limit set up by the financial services company.
13. High dimensional: data with many features or variables observed.
14. Labels data: is the data used to evaluate traders' trading performance. This included profit and loss, performance bonus, contribution per lot and hard stop counts.
15. Levene Tests: a statistical test to evaluate the equality of variances for two or more groups.
16. Mnemonics: characters used to pseudonymize the data.
17. Performance bonus: financial reward earned by the trader (in addition to monthly salary) based on the profit they make with their trades.
18. Predictions: output from a machine learning model about how a trader might perform in the future.
19. Profit or loss: money the trader makes or loses based on the trades they execute.
20. Recruitment stages: five stages of the recruitment process for recruiting traders. They are as follows: application forms, questionnaires, math tests, videos submitted by candidates and assessment day.
21. Stationarity: a property of time series data with certain statistical properties, such as their mean and variance do not vary across time.
22. SVM type of penalty used (L1 or L2): types of regularization techniques used to improve the generalizability of the support vector machine algorithm in machine learning.
23. Tech Savvy: individuals who are proficient in the use of modern technology, like computers and smartphones and have some idea of how AI systems depend on the data.
24. Volatility vs liquid market preferences: trading in products that have high variation compared to products with low variation in prices. Traders can usually exit their positions from the latter with a minimal

cost.

25. Volume and held positions: number of total futures traded for different products across multiple trades by a trader.

# Chapter 2: Phase 1 Research Background: Transparency, Artificial Intelligence and Education

## 2.1 Introduction

Transparency in AI, also referred to as ‘AI Transparency’ is a concept at its infancy and will develop over a period of time (Theodorou et al, 2017). The goal of this chapter is not to offer a ‘correct’ definition of transparency in AI, but to open-up the black box of *understanding transparency* by identifying the different interpretations of its characteristics that may be made from all the important terms in its definition. I also conceptualize how the meaning of transparency evolves with the different categories of AI in education stakeholders, and how these interpretations can be mapped to the resources required to make an AI system Transparent for those stakeholders.

Any definition of transparency in AI systems is bound to have some overlap with other ethical AI dimensions like explainability, accountability, fairness, privacy and safety due to its multifaceted nature. The broadest (but to a certain extent unsatisfactory) definition of Transparency in AI refers to a *process* with which the *information about an AI system is shared* with the stakeholders and this shared information *enhances the understanding* of these stakeholders. This means the AI powered ed-tech users’ existing background and understanding of AI can play an important role in determining the transparency of an AI tool. All the important terms within this definition can be interpreted in several ways:

1. *“Process”*: may be interpreted as a continuous phenomenon that is not dependent on a particular horizon of time, particular set of tools or a particular stage of an AI development pipeline. It can be open-ended or have its limits determined by that process’ creator.
2. *“Information”*: may be interpreted as all the details including decisions, decision-making processes, choices and assumptions made during the AI development process. It can be every detail of the AI development pipeline or just a particular aspect of it, irrespective of whether the

companies building AI products perceive that information to be useful or not useful.

3. “*Shared*”: may be interpreted as enabling someone else to know how an AI system is being built. It can take many different forms, from open sourcing the code of an AI system to providing documentation of how everything works or conducting training sessions with product experts to explain the details of an AI system.
4. “*Enhances the understanding*”: may be interpreted as increasing a person’s knowledge of a particular AI system or enabling them to fully understand every tiny detail of a particular system. It is a form of condition which ensures that to fulfill the requirements of transparency, whatever *information* is being *shared* must improve the understanding of the person with whom it is being shared.

These four elements of the definition broadly cover what may be referred to as Transparency of AI systems. These elements can also be taken as variables in determining the Transparency of an AI system. The variability in these four constructs adds to the complexity of Transparency in AI and blurs the boundaries between Transparency and other dimensions of ethical AI. I will now dig deeper into each of these components of transparency to conceptualize its variability in the context of education.

## **2.2 The variability of a ‘Process’**

The AI product development lifecycle starts from initial discussions on the purpose of an AI tool, its deployment strategy and the decision to start collecting data. It is a continuous process and does not end after the tool is deployed in the real-world. Transparency can be considered an integral part of this whole continuous process. It can be limited to the analysis or limitations of collected data (Yanisky-Ravid et al, 2019), algorithmic transparency (Garfinkel et al, 2017) of machine learning models being used or the deployment strategy and tools used for launching the product (Bhatt et al, 2020).

Felzmann et al (2020) have proposed a framework for transparency through the design of AI systems. The purpose of their framework is to bridge the gap between high level AI ethics principles and AI practitioners who are developing AI tools. One of the principles in their framework focuses on transparency as an integrative process throughout the AI tool's development pipeline. For Felzmann et al (2020), Transparency in AI has to take account of the entire AI development pipeline including the iterations for improvement after the tool is deployed in real-world.

Considering the complexities of an AI development pipeline, researchers have produced a number of frameworks to document the different stages of the AI tool development process. For example, Gebru et al (2018) introduced 'datasheets for data sets' to standardize the documentation of datasets (including their strengths and weaknesses) used to train machine learning algorithms. Mitchell et al (2019) have introduced model cards to document the strengths and weaknesses of machine learning models used to make predictions. Arnold et al (2019), from IBM have produced Factsheets to facilitate AI service providers in documenting their products' functionalities, performance, safety and security. All these proposed frameworks cover different aspects of the Machine Learning development pipeline or different parts of the AI tool development process. For example, datasheets from Gebru et al (2018) focus on documenting or making transparent only certain aspects of the data processing stage in the AI development process. In this context, the *process* in the definition of transparency is limited to the data processing stage.

This means a 'process' in the context of Transparency in AI can be interpreted as being instantaneous and limited to a particular stage of the AI development pipeline, or it can be continuous and encompass the entire AI product development and improvements lifecycle. Datasheets for datasets focus on the data processing stage of the AI tool development pipeline and model cards focus on the machine learning modelling stage of the AI tool development process. On the other hand, Fact Sheets cover a broader aspect

of the AI tool development process, focusing on tool's performance in general regarding different ethical AI dimensions.

### **2.3 The variety of 'Information'**

The term 'information' may be interpreted in many different forms in the context of Transparency in AI. It highlights 'what' is being made Transparent, or what 'information' is being shared with the stakeholders. It is also dependent on how the AI practitioners who develop AI products perceive the interpretation of 'process' above. If it is taken as limited to certain stages of the AI development pipeline, then only that information will be shared. If it is taken as a continuous process, then some details of the entire AI development pipeline may be shared with the stakeholders.

The problem with the second approach or complete transparency is that it can potentially lead to information overload for stakeholders (Eppler and Mengis, 2004) or the transparency paradox (Richards and King, 2013). Sharing everything with the stakeholders can potentially confuse them and make it more difficult for them to find the relevant information (Stohl et al, 2016). Some researchers like Heald (2006) have used the term 'transparency illusion' to illustrate this phenomenon.

It can be argued that on many occasions, AI products' stakeholders may not even know what information they need. What is useful for them or what kind of an impact a lack of transparency can have on them (Bogina et al, 2021). This is especially the case for stakeholders who are not tech experts and do not know exactly what kind of information from the entire AI tool's development pipeline will be useful for them.

The companies developing AI-powered products decide not only 'what' information is shared with the stakeholders, but also 'how' and 'when' it is being shared.

### **2.4 The interpretations of 'Shared'**



'Shared' in the context of Transparency in AI focuses on 'how' and 'when' the information is being shared with the stakeholders. Some might argue that steps like making the code of AI implementations public through GitHub or other tools are not very helpful for general public or stakeholders with no tech background. But it can have other positive effects. For example, AI engineers and practitioners know their work (code) will be viewable by the public in future, which reinstates the need to work towards public good (Elster; 1998; Chambers, 2004; Chambers, 2005; Naurin, 2007). It also means that practitioners know they can be held accountable for their work and will be answerable for the decisions taken and assumptions made in the development process.

Regarding the effectiveness of 'what' information is *shared* with shareholders, the dimensions of time, 'when', and form, 'how' it is being *shared* is also very important. For some AI systems, making the details of AI product development lifecycle public (like open sourcing the code or publishing freely accessible research papers) to be accessed any time might be sufficient. For example, for autopilots in cars, users do not need to know the details of how the car's autopilot is making every decision on the screen while driving a car. Separate documentation on how the autopilot works or its image recognition system is trained can be shared with the stakeholders to go through if they are interested.

For other AI systems, it might be more effective to share the relevant details exactly at the moment when decisions are being made. For example, for an AI-powered recruitment tool, if a candidate belongs to an under-represented group that is not presented in the training data, recruitment managers should be informed immediately that tool's predictions regarding this candidate are less likely to be accurate. In this context the information needs to be presented at the time of decision-making.

The dimensions of *time* and *form* can be very useful in determining the effectiveness of useful information being shared about AI systems. Irrelevant

information at the right time or useful information at the wrong time can lead to confusion and information overload as discussed above.

Some researchers have also argued against complete transparency such as Zarski (2016), Lepri et al (2017), De Laat (2018) and Carabantes (2019). Complete transparency (in which all the information about AI development is fully shared) like making the code of an AI tool public can hinder innovation. For example, for complete transparency companies may be sharing the secret sauce that makes AI work in certain contexts and provides them competitive edge over others.

## **2.5 The variety of meanings of ‘Enhancing the Understanding’**

This is one of the difficult requirements of transparency that makes the transparency of AI systems harder to achieve. It is partly because this part of the definition makes transparency dependent on the stakeholders of an AI system. It gives stakeholders or users the authority to decide whether an AI system is ‘enhancing their understanding’ or is transparent.

Transparency’s capability to ‘enhance the understanding’ of an AI system for its stakeholders can be taken as more of a condition for an AI system to be considered as transparent. It shows that transparency is inherently dependent on the people for whom it is targeted. It also means that a construct that might be considered transparent for one person might be a black box for another, if it is not enhancing the other person’s understanding of how an AI system works.

According to Turilli and Floridi (2009) transparency is dependent on factors such as the availability of information, conditions of its accessibility and how this transparent information may pragmatically assist in decision-making. For an AI system to be transparent, there must be a right mix of the quality and quantity of information shared, accessibility of that information and if it facilitates decision-making. ‘Enhancing the understanding’ of a stakeholder is dependent on the background knowledge and the context in which an AI

system is impacting that stakeholder. This means that the kind of AI tool being built, and its stakeholders determine the effort required by AI companies and practitioners to make their tools transparent. Making an AI tool transparent for AI engineers might require considerably less effort in making it transparent compared to an AI tool for general public who are not tech savvy (Tamboli, 2019).

Based on the various interpretations of *process, information and shared* mentioned above, 'enhancing the understanding' can be related to a certain stage of the AI development process or for the entire development pipeline. Theoretically, this condition of making an AI tool transparent refutes the issue of cognitive load or information overload mentioned in section 2.3. It can ensure that the information being shared with the stakeholders is serving its purpose rather than causing confusion.

In enhancing the understanding of stakeholders, it is important to make a distinction between transparency in machine learning powered AI systems and rule-based AI systems. There has been some study of the transparency of rule-based AI in education, focusing on specific ed-tech products like open learner models (Bull and Kay, 2007; Bodily et al, 2018; Bull, S., 2020) and intelligent tutoring systems (Polson and Richardson, 2013; Mousavinasab et al, 2021) to enhance the understanding of stakeholders. But this work does not take account of machine learning powered AI's capabilities and their impact on education, like predicting grades with neural networks (Zhang et al, 2021), clustering learners based on their performance (Sari et al, 2021) or algorithmic decision making for students' admissions. In rule-based systems, the risks of AI going completely wrong are comparatively lower as AI's decisions are limited by pre-defined rules and outcomes (Marcus, 2020). The focus of such systems is mostly on data visualizations and analytics based on the real-time data that is being collected (Mangaroska and Giannakos, 2018) rather than machine learning powered AI systems which are the focus of this thesis.

## **2.6 Transparency and other ethical AI dimensions**

In this section, I explore how these varying interpretations of transparency affect its positioning within the wider context of ethical AI. Figure 4 prepared during this research, highlights one of the ways in which various dimensions of ethical AI may relate to each other. Considering there is no universality in the definitions of these terms and each of them has been interpreted in various ways (such as transparency shown above), there is no definite answer to how these dimensions overlap (Hao and Stray, 2019; Prunk and Whistestone, 2020; Anjaria, 2021; Brendel et al, 2021). Some of the missing ethical AI dimensions from the figure like privacy can be covered under safety, benevolence can be covered under fairness and non-maleficence can be covered under accountability.

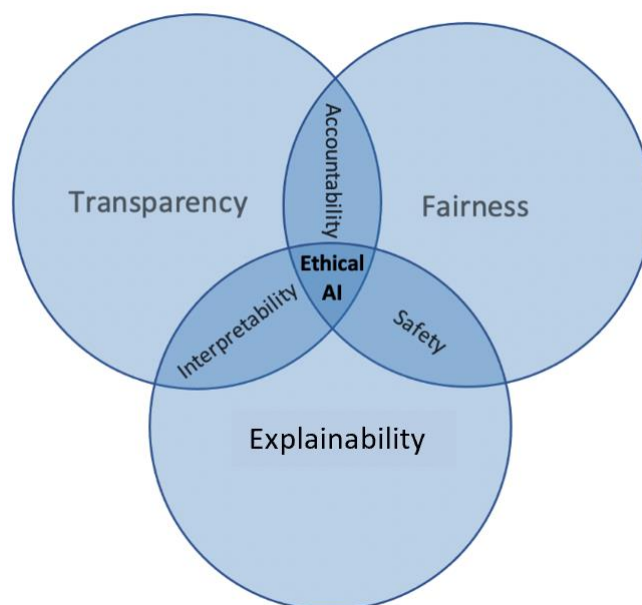


Figure 4: Potential overlap between Transparency and other ethical AI dimensions

Figure 4 shows one of the ways in which the different dimensions of ethical AI may overlap. There is no definite answer to how one dimension relates to another. This relationship between different dimensions depends on how each dimension is perceived and how the lack of a particular dimension may relate to some form of bias in an AI system (Baker and Hawn, 2021). For example, according to the figure above explainability and fairness together can lead to AI safety. Explainability assumes that users understand how an AI system is

making decisions and Fairness ensures that the system does not discriminate against any group. But, considering the overlap of the varying interpretations of these dimensions of ethical AI, it may be argued that transparency and interpretability are also necessary conditions to ensure the safety of an AI powered ed-tech (McDermid et al, 2021).

### **2.6.1 Explainability and Interpretability**

There are a number of tools and frameworks available to make AI systems explainable for stakeholders (Roscher et al, 2020; Holzinger et al, 2018; Bhatt et al, 2020; Samek et al, 2017; Bracke et al, 2019; Spinner et al, 2019; Ahmad et al, 2018). But, if the stakeholders cannot understand those explanations, they are not of much use. There have been a number of studies on formulating explainable and interpretable AI systems that are effective (Samek et al, 2017; Doshi-velez et al, 2017; Lipton, 2018; Gilpin et al, 2018) but many times they do not clearly specify the goals of explainable AI. These goals are dependent on stakeholders who will be using the AI system and for whom the explanations are being derived.

To make the explanations of an AI system understandable for stakeholders like educators, instructors, principals and ed-tech experts, different tiers of users can be inducted in an AI system dependent on the background knowledge and technical acumen of the stakeholders. If the explanations of an AI system are not enhancing the understanding of stakeholders regarding how an AI system works, it cannot be considered transparent. On the other hand, if the explanations are transparent and enhance the understanding of stakeholders, they can be considered interpretable as well.

The terms *interpretability* and *explainability* when applied to AI systems are at times used interchangeably to illustrate how easy it is for humans to understand the cause of a particular decision by an AI system (Linardatos et al, 2021). Doshi-Velez and Kim (2017) define interpretable AI systems as the ones that 'present (their decisions) in understandable terms to a human'. Mittelstadt et al (2019) have discussed the different types of explanations that can be produced by black-box AI systems. When a human interacting with an

explainable AI system can fully understand the explanations offered by the AI system, we call this AI system interpretable. An explainable system is not necessarily interpretable for everyone, but every interpretable system must be explainable. Hence, in the context of transparency, my focus is on interpretable systems where stakeholders (based on which transparency tier they are in) can understand the reasons behind AI's decisions.

In the context of explainability and interpretability in different educational settings, the requirements of transparency can be broadly divided into two parts: firstly, sharing the information with AI system's users, and secondly, how is that information shared and if it enhances the understanding of stakeholders. The first part focuses on explainability and second part focuses on interpretability. The first half of this definition focusing on sharing of information makes transparency a necessary condition to make an AI system interpretable, and second half of the definition focusing on enhancing users' understanding makes interpretability a necessary condition for transparency. With these interpretations of transparency and interpretability we cannot have one, without the other. An AI system must be interpretable to become transparent, and it needs to be transparent to become interpretable.

### **2.6.2 Fairness**

The urgency towards AI development and deployment has repeatedly cost users in the form of fairness or some form of discrimination against particular groups (Whittlestone et al, 2019). Fairness in AI systems, like other dimensions of ethical AI is a challenge for AI practitioners and the companies building these tools as there is no universally agreed upon definition of fairness in AI. Several researchers have presented the taxonomies and toolkits for defining and computing fairness in different contexts (Barocas et al., 2019; Caton & Haas, 2020; Kizilcec & Lee, 2020; Mehrabi et al., 2019; Mitchell et al., 2021; Verma & Rubin, 2018). Chouldechova (2017) has discussed how different stakeholders may have different notions of fairness. Interestingly, some researchers like Karen Hao and Jonathan Stray (2019) have concluded that to decide whether an algorithm is fair or not also depends on how we define fairness or bias. They have shown how there

needs to be different benchmarks of fairness for two different groups of people to decide if an individual from a particular group will go to jail or not. Their analysis was based on the fairness of AI systems used in the courts in US. Even if a particular AI tool like COMPAS (Dieterich et al, 2016) used in the context of judicial decision-making in US is taken as fair, it would become biased with time as it will be trained by its own produced data that has been sending more individuals from a certain group to jail. Hence, the situation may potentially worsen with time. It means deploying a fair AI system in real world does not guarantee that it will remain fair after interactions with the real-world data.

There are no perfectly fair AI systems (Corebett-Davies and Goel, 2018). Irrespective of how the company building an AI-powered ed-tech tool interprets fairness, transparency can be considered a pre-requisite for fair AI systems. To make sure that the system is fair or performs as expected, AI practitioners need to firstly analyze the data and identify any biases (Mehrabi et al, 2021; Zhou et al, 2021). There will always be some form of bias in the data (Ntoutsi et al, 2020) but the details of this bias and its impact should be shared with the stakeholders. Similarly, the machine learning model chosen for predictions and tools used for deployment have their strengths and weaknesses. All these details should be shared with the stakeholders in an easily understandable manner to make sure that the system is fair. With this interpretation of fairness, the requirements for transparency seem to be the first step towards fair AI systems.

Madaio et al (2020) have presented a detailed checklist to ensure fairness in AI tools after thoroughly reviewing different stages of AI tool development processes by consulting 48 AI practitioners. Many of the checks and balances that the AI practitioners propose in different stages of an AI product lifecycle including envision, define, prototype, build, launch and evolve are also an integral part of making the AI development pipeline transparent and can be applied in educational contexts. To make sure that an AI system is fair or is not used in a manner that discriminates against certain groups of learners or educators, AI practitioners need to be transparent about how that AI system is

built. What are the decisions, choices or assumptions made in the development process? In which context should this AI system not be used? And how do they interpret the fairness of their AI system?

Every AI tool built has its strengths and weaknesses. It works as expected in certain contexts and does not perform as good in others. To make sure that the AI system in an ed-tech is fair and is used as intended, AI practitioners need to be transparent about these strengths and weaknesses and share how their decisions and assumptions can impact the performance of the tool in different contexts.

### **2.6.3 Accountability**

Accountability in AI has been an ongoing challenge between companies developing AI tools, stakeholders impacted by these tools and regulators responsible for protecting the rights of citizens. It is based on a simple legal principle that if anything goes wrong, there has to be someone responsible for it. Mishaps and unintended consequences of AI systems are not uncommon (Osoba and Welser, 2017; Mujtaba and Mahapatra, 2019; Challen et al, 2019; Johnson et al, 2019). These unintended consequences of AI systems can have severe long-term effects on their users, for example a recruitment tool rejecting a particular candidate only because they are female (Ahmed et al, 2018), an admissions tool wrongly predicting low grades for a student or a legal tool sending offenders to jail only because they belong to a particular minority group (Re and Solow-Niederman, 2019; Završnik, 2021) can have serious mental and financial repercussions. Hence, accountability of the companies developing these AI tools is of utmost importance.

Accountability is especially needed when an AI system in education is unfair. But to identify if a particular AI system is unfair for a certain group of learners or educators, or why it is not fair in certain contexts, transparency can be important. For autonomous AI systems, companies building the AI-powered ed-tech tools may be held responsible for any mishaps or malfunctioning (Raji et al, 2021). But, within an ed-tech there can be tens, hundreds and at times even thousands of engineers, AI practitioners, domain experts and software



developers working on the development of an AI tool. This can make communication within teams and decision-making complex and time consuming. It also makes the identification of mistakes and personnel responsible for it very difficult (Kim et al, 2020).

Transparency may be interpreted as a necessary condition for accountability. If an AI system malfunctions and a judge in the court of law needs to convict the person responsible for tool's unfairness, they need to first understand how the tool works, what it lacked, was it a careless mistake or an intentional choice and who was responsible for it. Some form of transparency seems to be necessary to enhance the understanding of judges and ensure accountability in this context.

It is necessary to have Transparency implemented in every stage of the AI tool development process to document all the decisions, choices and assumptions made in the entire process. This makes the diagnosis of the tool, identification of the exact cause of malfunctioning and which team or individual is accountable for it, much easier.

Many AI tools aim to empower humans to take the final decision (Jotterand and Mosco, 2020). In such scenarios, transparency is extremely important to make sure that 'human in the loop' understands how the AI tool is making a decision, when to trust its judgement and when to ignore its predictions or recommendations. In such scenarios, the decision-making authority is always in the hands of 'human in the loop', but they need to have a thorough understanding of the tool. This is possible if the entire AI tool development pipeline from data collection to final deployment is transparent for the 'human in the loop'. For example, if a company buys an AI-powered recruitment tool to enable its recruitment managers to make more informed decisions, it should be the company's responsibility to offer training to those office managers on how to use the tool, what are its strengths and weaknesses and in which contexts the tool should not be used at all.

This shows that irrespective of an AI tool being completely autonomous or having a 'human in the loop', and for all the stakeholders involved in the AI tool development process, transparency is the first step towards accountability. If the stakeholders do not have enough knowledge about pros and cons of every decision taken, strengths and weaknesses of third-party services used and assumptions made throughout the development process, then they may not be held responsible for mishaps. Someone in the company developing the AI tool must make sure that the entire process is transparent, so mistakes can be easily diagnosed and accounted for.

#### **2.6.4 Safety**

AI safety is another very important dimension of ethical AI that aims to ensure that AI systems add value to decision making processes and are used as intended (Yampolskiy and Spellchecker; 2016; Irving et al, 2018; Irving and Askill, 2019). It is dependent on the decisions taken throughout the AI development and deployment process which can range from a few weeks to years. To keep account of the decisions taken, assumptions made and experiments conducted in the AI development process of an ed-tech product, it is very important to be transparent about them and document these details. From this interpretation, the methodologies of ensuring AI transparency and safety seem to be directly correlated. They are a part of the design methodology, starting when the AI systems are being planned and continue even after the systems are deployed in real-world.

Ortega and Maini (2018) from DeepMind's safety research team have presented a framework that divides the safety of an AI system into three parts, Specification, Robustness and Assurance, as shown in figure 5 below:

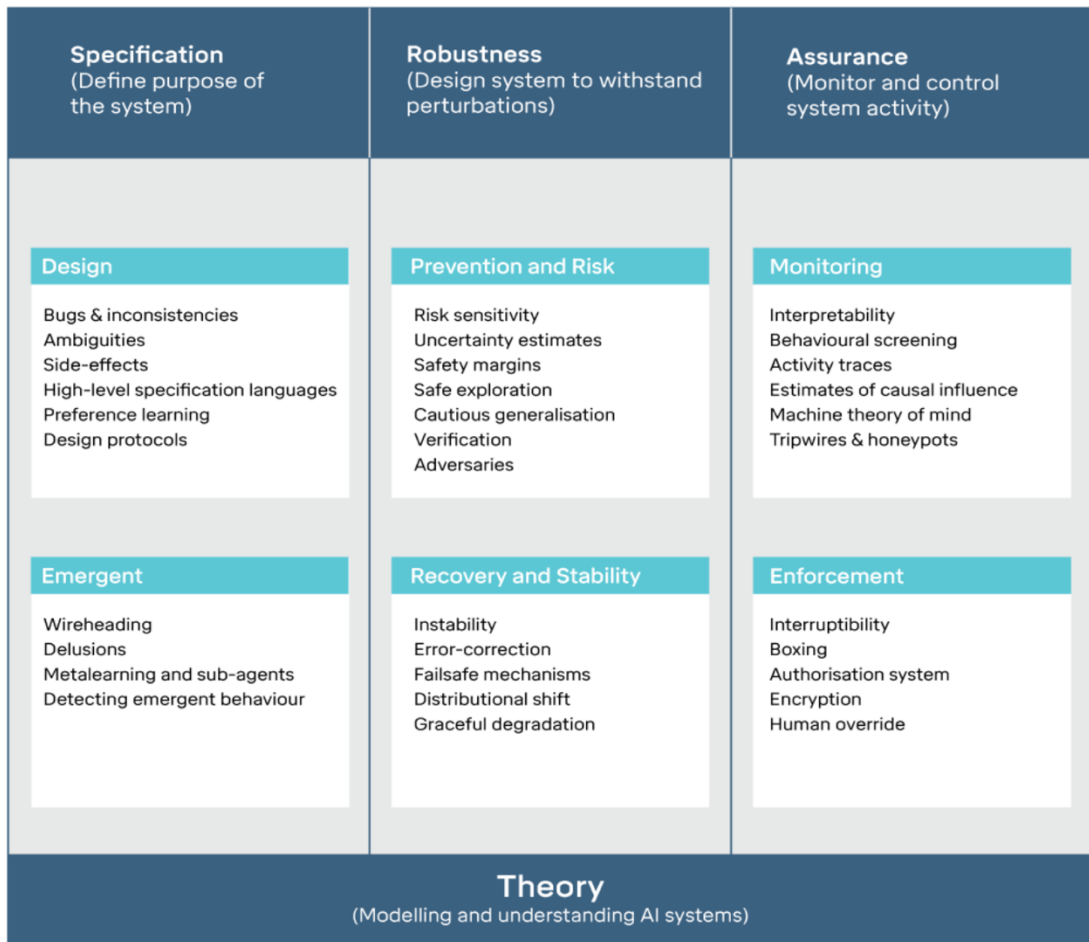


Figure 5: Building Safe Artificial Intelligence: Specification, Robustness and Assurance

Specification refers to the goals of the system. Three types of specifications of an AI system in the context of an AI-powered ed-tech tool can be explained as follows:

- Ideal Specification:** this refers to the best possible results expected from an AI system with no bias as discussed in section 1.3. For example, in an ed-tech tool screening candidates for admissions, this system would admit the best performing candidates through accurate predictions without any kind of bias.
- Design Specification:** this defines how the AI system is built. In an ed-tech tool this may involve the data collection process, feature engineering and machine learning algorithms used to allocate

applicants to a cluster. It also includes the graphical user interface with which AIED stakeholders may interact with an AI system.

- **Revealed Specification:** this is the actual result of the AI system, irrespective of ideal specification and dependent on the design specification. In an ed-tech tool, this may be in the form of an AI system admitting the best candidates to a university but being biased against gender<sup>10</sup> or a certain group<sup>11</sup> or a race.

For safe and transparent AI development in education, documentation of all three specifications above is very important along with all the steps taken to ensure robustness and assurance of AI tools, as shown in figure 5 above. Transparency of an AI-powered ed-tech product should consider all the measures taken for ethical AI development. It should particularly focus on the specification of an AI system, and answer questions like in which contexts the system works at its best or when is it better to avoid using an AI system.

The identification of an ideal context and specification during which an AI system performs at its optimum also helps in identifying the diversity and inclusion limitations of an AI system. This has been a long-lasting challenge for AI community (Fosch-Villaronga and Poulsen, 2022; Chan et al, 2021; Knox et al, 2019). Transparency as conceptualised in this research, can play a very important role in enabling more diverse and inclusive AI-powered ed-tech products by highlighting these limitations of AI systems and creating awareness about them.

Figure 6 highlights the importance of transparency in pointing out the exact differences between Revealed and Ideal specifications (like diversity and inclusion limitations) of AI systems: what the AI system can actually do in real

---

<sup>10</sup> <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10?r=US&IR=T>

<sup>11</sup> <https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1>

world scenario and what the end-users might think the AI system is ideally capable of doing in any scenario.

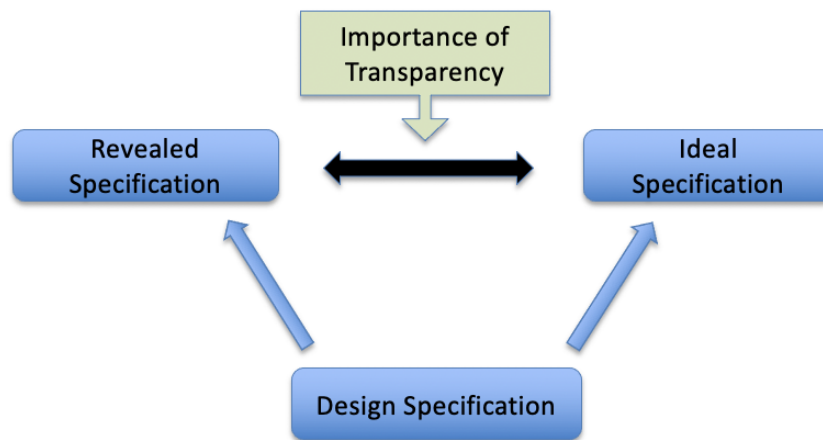


Figure 6: The importance of transparency in bridging the gap between Ideal and Revealed specifications of an AI system

The role of transparency is to reveal the gap between an ideal and a revealed specification by making the machine learning models and results of accuracy metrics in different contexts explicit and understandable for the stakeholders. For example, the importance of gender, age or location of candidates when applying for jobs. This would show the stakeholders when to trust an AI system and when to avoid it.

After ensuring the robustness and safety of AI tools in certain conditions and scenarios like in classroom, lab or a virtual learning environment, it is still very important to have measurement tools and protocols in place to assess the quality of decisions being made by an AI system (Doshi-Velez and Kim, 2016). This is where Assurance from figure 5 comes in. It ensures checks and balances on the performance of an AI tool in real world that can have unlimited scenarios and contexts. Amodei *et al.* (2016) illustrate this point with an example of an AI-powered cleaning robot that is responsible for cleaning offices. If this robot has never seen a pet before (or never trained on a dataset that has pets in pictures) and is deployed in a pet-friendly office, it can lead to unexpected consequences. It is very important for AI companies to provide guidelines to users on how to use their products and when not to use their products. Many times, companies may avoid sharing such information with

the users of AI products because it may contradict with their brand messaging as the contexts in which a typical AI product would not work can usually be a lot more than the contexts in which it would perform as expected.

AI systems are usually built with a particular dataset that is considered a (partial) representation of the real world's educational setting, a machine learning model that has its own strengths and weaknesses and a deployment strategy using certain tools that can impact the performance and outlook of an AI tool. With the complexities of dynamics created by learners and teachers, many things in any of these three different stages can go wrong. To ensure the robustness of these systems, AI practitioners usually take a number of measures to determine the safe limits of an AI system by evaluating the tool's performance in unexpected scenarios. In this case, it is very important for the companies developing these AI systems to document the robustness measures they take throughout the AI development pipeline. This process begins from the initial discussions on the problem that an AI system will solve, followed by the data collection and tool development and deployment strategies as discussed below.

## **2.7 The Artificial Intelligence Development Pipeline**

An AI development pipeline is an integral part of the AI tool formation and is a continuous process that starts from the decision to build an AI tool and continues even after the AI tool is deployed in the world. It can take weeks, months or at times even longer to build impactful AI products. This makes the AI tool development process complicated, resource intensive, complex and time-consuming. The discussion on transparency of AI tools would be incomplete without an in-depth exploration of the AI tool development process.

A great deal of work has been done to identify the different stages in the process of AI tool development and their importance (Amershi et al, 2019). For AI projects, practical work usually begins from the decision to collect data and continues until the ML model is deployed in production (Studer et al, 2021). This is followed by iterations to improve the machine learning model's

performance. It is a step-by-step process involving data cleaning, then feature engineering, model selection, model training, model deployment and eventually model improvements through further iterations. All these steps in the AI development process can be impacted by the ed-tech company's decision to make them transparent (Larsson and Heintz, 2020) and transparency considerations need to be ingrained in the entire AI development process.

Microsoft Azure conceptualises the AI development process through three stages on which the ML pipeline needs to focus<sup>12</sup>:

- **Stage 1:** Data preparation including importing, validating and cleaning, munging and transformation, normalization, feature extraction and staging.
- **Stage 2:** Training configurations including model selection, training and validating. This may include making choices regarding hardware compute resources, distributed processing and progress monitoring.
- **Stage 3:** Deployment, including versioning, scaling, provisioning and validating model efficiently to ensure that the model works as intended for its users.

## 2.8 An AI Development Pipeline in Education

An AI development pipeline in education can broadly be divided into three categories: data collection and processing through real-world classroom settings or virtual learning environments, machine learning modelling and evaluation with algorithms like decision tree (Agarwal et al, 2012), neural nets (Hu, 2017) or support vector machines (Zhou et al, 2010) and the deployment of an AI system in the real-world through cloud technologies or local servers (Halde, 2016; Kucak et al, 2018). Matrinez et al (2019) have presented a canonical architecture for machine learning development applied to human language technologies. Following the same principles, we can divide the AI

---

<sup>12</sup> <https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-ml-pipelines>

tool development process in education in terms of transparency into three stages, as shown in figure 7:

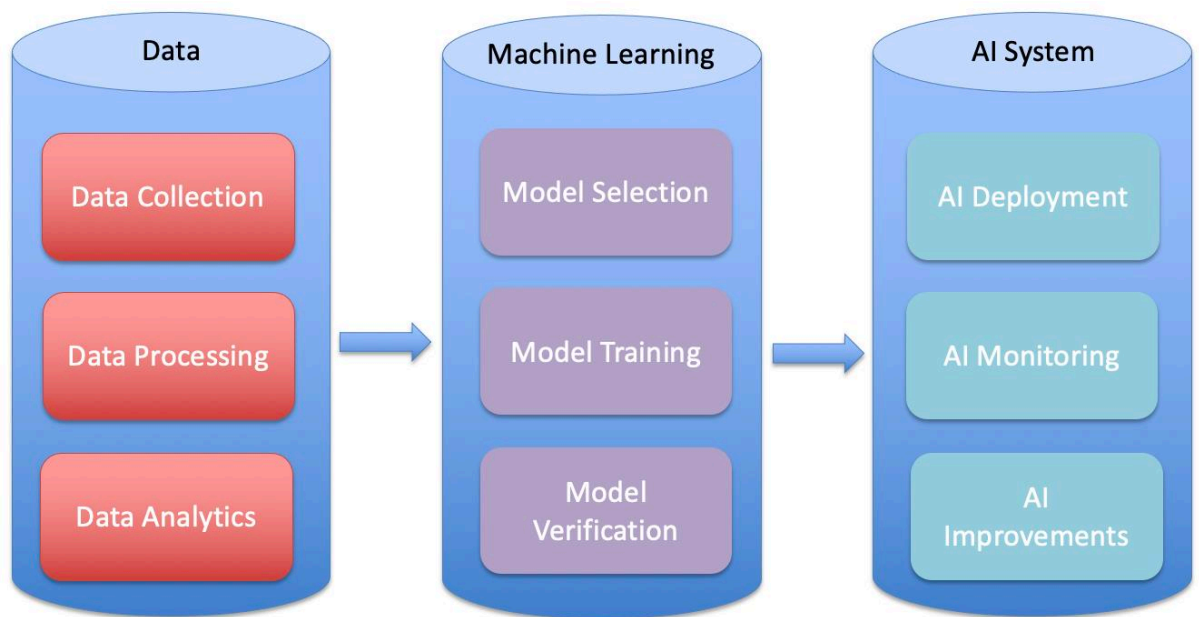


Figure 7: The canonical architecture for AI tool development in terms of transparency

Irrespective of which approach is being used to develop AI systems, various measures can be taken by the ed-tech companies to develop these systems ethically. Every component within each canon requires individual attention, resource allocation, time commitment and a different set of development tools. Each of these components within the canons needs to be addressed independently to make them transparent for stakeholders. These unique requirements of each sub-component within each canon in fig 7 highlight the complexity and range of skills required for developing an AI system.

### 2.8.1 Planning

The AI tool development process starts from planning, and so does the need for transparency in the planning process. There are two main questions that need to be considered when planning the development of an AI product. These questions help in determining how the tool will be used, what types of users will be using the tool, how the tool can impact the lives of the stakeholders involved and what measures need to be taken to make the AI



development process transparent for stakeholders. For ed-tech companies for whom ethical considerations, such as transparency should be a priority, there is also a third point that needs to be considered as discussed below:

#### 1. Impact on Stakeholders:

The first point to be considered when planning the development of an AI tool is the kind of impact it can have on stakeholders (AM Cox, 2021; Leahy et al, 2019), especially the users who will be impacted by the tool's decisions like learners or educators. Some AI tools like a product recommender system on an e-commerce website or a system that shows ads to users while surfing the web, may not usually have a major impact on users' lives unless they are within a certain domain, such as gambling. These AI tools are low-risk compared to AI tools that can have a direct life-changing impact on their users. For example, an ed-tech tool that determines whether a candidate should be admitted to a university or not, can have a huge impact on the candidate's life. Similarly, AI tools have been used to process loan applications for candidates (Demajo et al, 2020) and taking several healthcare decisions (Yu et al, 2018). AI tools that make such decisions can potentially have major long-term implications for their users.

#### 2. Human in the Loop:

The second point to be considered when strategizing the development of an AI tool is whether we want the tool to be completely autonomous with a 'human out of the loop' approach, or whether we want a human to supervise the working of the AI tool as a 'human on the loop' or if we want the final decision-making authority to stay with humans as 'human in the loop' (Zetsche et al, 2020; Jotterand et al, 2020). This choice is dependent on how we perceive the potential risk of using a particular AI system or kind of impact it can have on the stakeholder as shown in the figure 8 below. This choice can be subjective and is usually determined by the kind of impact an AI tool's decisions can have on its users. If the impact of the tool is high, it can be considered a high-risk tool, and vice versa. AI systems like a recruitment tool, loan

applications processing tool, or a law enforcement tool can be considered high-risk because their decisions can have a huge impact on the lives of their users.

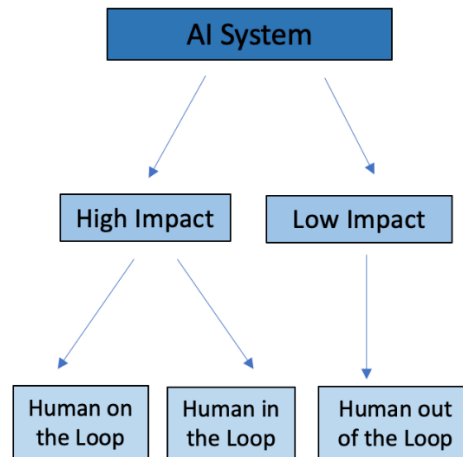


Figure 8: An overview of the role of human in using an AI product based on the product's impact

For low-risk AI applications, such as recommender systems, complete autonomy with 'no human in the loop' can work, because wrong decisions by the AI in such contexts will probably not have a big effect on someone's life. For high-risk AI applications, for example a recruitment tool or a grades prediction tool, completely relying on AI systems with no human involvement can be very risky. Wrong decisions by AI in such contexts can completely transform someone's life. For example, a drop out prediction tool wrongly predicting that a particular student will drop out in six months can potentially have an adverse psychological impact on the learner and effect how their teachers and parents treat them. In such applications, a 'human in the loop' approach is preferred to keep the final decision-making authority in the hands of humans. In this context, the final decision is always taken by a human and their own judgment is preferred. This ensures the accountability of AI systems and adds another layer of security when an AI system's decisions are completely off-track. The role of AI systems prepared with the 'human in the loop' approach is to empower the humans to make more informed decisions.

Ethical considerations produce a third factor that also needs to be considered when planning the development of an AI tool.

### 3. Background of stakeholders:

The background of people who will be using the tool or who will be impacted by the tool. This can take the form of the personification of these users in terms of their technical background and capacity to understand how different components of an AI system work, into personas. Knowing this in advance helps in planning the developmental strategy of the tool and in determining resource allocation. For example, if the users of an AI product are educators who do not have a technical background and are not very tech savvy, then additional developmental resources may be required to present the explanations of an AI system in an easily understandable manner for such users. On the other hand, if an AI system is targeted at AI practitioners or researchers with a strong technical background, then it might be a bit less resource intensive to make an AI tool's explanations understandable for them.

The three cannons of the AI tool development process with respect to transparency: Data, Machine Learning and AI system as shown in figure 7 are now discussed in detail with regards to their relationship to transparency in educational contexts.

### **2.8.2 Data Collection and Engineering**

Once the problem that the AI tool being developed is going to address has been identified and the development strategy is finalized, data collection begins. Data collection is one of the most important stages in the AI development process and assumptions made during this time can have long-term effects on the performance of the tool. There are several ways in which the relevant data for the problem at hand can be collected (Sapsford and Jupp, 1996; Gallagher, 2009). The company or the team building the AI tool can collect the data from real world themselves, called primary data collection (Roos et al, 1987; Ibert et al, 2001) or they can use the data already collected

by others, called secondary data collection (Daas and Arends-Toth, 2012; Hox and Boeije).

Different data collection methods have their pros and cons that can be made transparent for stakeholders (Jars-Quant, 2018). Primary data collection empowers the AI team to collect any kind of data they want. The AI practitioners can define the proxies for the metrics they want to measure and start collecting data accordingly. Primary data collection methods can be quantitative through surveys and questionnaires (Parajuli, 2004) and/or qualitative through interviews and observations (Hayman et al, 2012, Jain, 2021). They can be collected from focus groups or sampled from public (Cyr, 2016). These collection methods are expensive, time consuming and require manual labor. Minnaar and Heystek (2013) evaluated the importance of online surveys as a primary data collection method in educational research and highlighted the challenges of this technique in collecting sufficient data. Primary data collection methods enable companies to control the distribution of different sensitive variables in the data. For example, the company collecting the data can ensure that have equal distribution of males and females, from a certain age group and geographic region and from different ethnic groups. It empowers the companies to collect relevant data for their AI tool and mitigate any deficiencies in this initial stage.

On the other hand, secondary data techniques are more cost effective and less time consuming to acquire because they have already been compiled by some third party and are readily available (Johnston, 2017). But this also means that they may not perfectly match the requirements of the AI team who is building the tool. Secondary data would have been mostly collected for another purpose and may not have any documentation available on why and how it was collected, what were its strengths and weaknesses or how it could be improved (Smith et al, 2001). These considerations should be taken into account to make the data selection and collection process transparent. Thomas et al (2001) have shown how the data collected from sampling of students in higher education can be difficult to utilize in different contexts. Burchinal et al (2016) have used secondary data from eight large studies of

pre-school children to show that children benefit the most from early care and education when certain quality thresholds are met and/or it is of longer duration.

These primary and secondary data collection techniques can be used to build/train the AI systems. But, for evaluating the impact and usefulness of a particular AI system, only primary data regarding the usability of a particular AI system can be used.

Once data collection is complete, then the data is cleaned, processed, engineered and analyzed to get it in the right form for the machine learning modelling stage. A number of tools and frameworks can be utilized for completing these tasks (Katal et al, 2013; Rao et al, 2019; Redi et al, 2021; He and Garcia, 2009; Wu et al, 2013; Romero and Ventura, 2013). All the measures taken to handle the data and the tools used to store, engineer, and analyze it can be documented for transparency considerations.

The two main approaches for analysing the data and identifying any correlations or causations are frequentist and bayesian. There are some fundamental differences between bayesian and frequentist statistical approaches. While frequentists analyse probabilities through the frequency of repeated events, bayesians focus on the probable uncertainty in the data. Frequentists keep the model parameters fixed but bayesians change the model parameters as they are conditioned on the present data (Cox, 2006; Senn, 2003; Wagenmakers et al, 2008).

Bayesian statistics has gained a lot of popularity among researchers and practitioners of artificial intelligence. This is for two main reasons: firstly, a number of machine learning applications failed in the real world and secondly, there was too much focus on adopting frequentist (Mayo and Cox, 2006) statistical approaches in feature engineering of raw data and machine learning models. Bayesian theorem's ability to take account of uncertainty in the data makes it a lot more suitable for real world applications, compared to frequentist statistical approaches.

As shown in figure 1, there will always be a difference between the data that is sampled from the real world and how human experiences perceive it. The digital data reflects only on a certain proportion of the real-world. This discrepancy between the data and real-world needs to be identified in the data processing stage of the AI development process (Ntoutsis et al, 2020). Carvalho et al (2019) have provided a systematic literature review of how researchers in the past have used different proxies for data collection to solve the problems at hand. In education proxies have been commonly used to track learner progress, evaluate learning outcomes and explore the efficacy of learning content. Cukurova *et al* (2018) have shown through the NISPI (Nonverbal Indexes of Students' Physical Interactivity) framework how school and university students' hand positions and head direction data can be utilised to judge students collaborative problem-solving competence. Alwahaby *et al* (2021) have provided a thorough literature review of ethical considerations in multimodal data for learning analytics.

Data that is collected from the real-world to train machine learning algorithms can never be perfect. It will always have deficiencies that need to be identified, mitigated and shared with the stakeholders to ensure ethical AI development (Jo and Gebru, 2020) in education.

The limitations of the data can be identified from the exploratory and statistical analysis using frequentist and bayesian techniques mentioned above. To mitigate these data limitations, several approaches have been identified by AI researchers and practitioners (Veale and Binns, 2017; Balayn et al, 2021; Marda, 2018). Abusitta *et al* (2019) have proposed Generative Adversarial Networks also known as GANs (Goodfellow, 2016; Creswell et al, 2018; Karras et al, 2019) for producing synthetic data to mitigate biases in AI systems. A number of researchers have proposed class balancing techniques in unbalanced datasets to enable comparatively fairer outcomes (Dal Pazzolo et al, 2010; Yan et al, 2020; Waheed et al, 2021). Thammasiri *et al*, (2014) have shown how these class balancing techniques can help in mitigating class imbalance in the context of predicting student retention where a lot of

students register for a particular class but only a minority drops out. Methods like synthetic data and class balancing techniques can help in mitigating the effects of data limitations, but these measures cannot guarantee complete removal of the deficiency of real-world data (Lum, 2017).

The decisions and assumptions made during the data processing stage of an AI development process that affect the performance of the tool need to be shared with the stakeholders in an easily understandable manner. This information may not necessarily be a part of the ed-tech product itself but can be a part of the training material when educators are on boarded. The role of transparency and how it is perceived by the ed-tech companies can play an important role in decisions about what information is shared with the stakeholders and how it is shared. For example, some ed-tech companies use research-based methodology and publish research papers in journals or academic conferences to illustrate the effectiveness of their AI-powered ed-tech. Others may open source the software code of their product for replication or share the details of their AI development process during the training of prospective users.

### **2.8.3 Machine Learning Modelling**

The Machine Learning (ML) modelling stage is the epicentre of AI systems. It plays a very important role on what type of predictions or decisions are made by AI systems, how they are generated and if they can be explainable for stakeholders. There are three main types of ML techniques (Jordan and Mitchell, 2015) that are commonly used in real world settings.

1. **Supervised Machine Learning:** This class is one of the most used class of algorithms that are trained on labelled datasets (Singh et al, 2016; Cunningham et al, 2008; Hastie et al, 2009). Such techniques need an output variable that the algorithm is supposed to predict in the training data from which they learn (Kotsiantis et al, 2007). These algorithms are especially effective for classification tasks where the input needs to be grouped into different classes based on its similarities and

differences with the datapoints of each class (Osisanwo et al, 2017; Mohamed, 2017; Bhavsar, 2012).

2. Unsupervised Machine Learning: This class of technique is mostly used to learn from raw unstructured data where the output variable is not used as labels (Usama et al, 2019; Khanum et al, 2015). These models learn the distribution of datasets from their structure and make decisions on new data based on these learnt distributions. Cam *et al* (2021) have shown how these techniques can be used to predict students' success rate in a course. Fwa and Marshall (2018) have shown how unsupervised learning can be utilised to model student engagement through head pose, keystrokes and action logs data from intelligent tutoring systems.
3. Reinforcement Machine Learning: This technique has become very popular among researchers in the last few years (Hao, 2019). It is the 'problem faced by an agent that learns behaviour through trial-and-error interactions in a dynamic environment' (Kaelbling et al, 1996). Reinforcement Learning has led to a number of research breakthroughs in AI especially in games like playing Atari at expert human-level (Mnih et al, 2013) and mastering chess, shogi and go (Silver et al, 2018). It has also contributed to solving the protein folding problem in healthcare (Callaway, 2020). Within education, reinforcement machine learning techniques have been used to model students' learning styles (Dorca et al, 2013), choosing pedagogical policies in intelligent tutoring systems (Iglesias et al, 2009) and building adaptable educational tools (Bennane, 2013; Iglesias, 2009; Shawky and Badawi, 2018).

There are several algorithms that can be utilised to achieve a particular task when it comes to the ML techniques mentioned above. In education, methods like neural networks have been utilised for predicting student course selection (Kardan et al, 2013), assessing quality in technical education (Mahapatra and Khan, 2007) and evaluating teaching quality (Hu, 2017). Decision trees have been used for educational data mining (Agarwal et a, 2012), modelling group decision-making in education (Chang and Wang, 2016), predicting student



drop-out (Quadri and Kalyankar, 2010) and predicting student satisfaction at a business study program at a private higher education institution (Skrbinjek and Dermol, 2019). Similarly, support vector machine (svm) algorithms have been used by Gil *et al* (2021) to predict first year students' academic success in higher education and Zhou *et al* (2010) have used svm algorithms to identify 'children's health and socio-economic determinants of education attainment'.

The choice of a ML technique or algorithm to be used in education is determined by three main factors or priorities that are setup by business leaders (Luan and Tsai, 2021) or AI researchers and practitioners working on the ed-tech. These factors are discussed in detail below:

1. The accuracy of the results, which should be as high as possible. This is quite intuitive for real-world scenarios and makes a strong business case as accuracy of AI tools can be a major selling point to clients. Yin *et al* (2019) have shown how accuracy effects trust of people on AI systems. They show a positive correlation between accuracy and trust. At times, business leaders may prefer a certain type of accuracy, for example they want fewer false negatives at the cost of more false positives in results. In such contexts the machine learning models need to be adjusted accordingly. In educational contexts the feedback loop for ed-tech companies to evaluate the accuracy of their AI systems can be difficult, but teachers can play a pivotal role in determining the accuracy of AI-powered ed-tech products (Zhao and Liu, 2018). For example, if an AI system recommends a piece of content to learners based on their past performance and learner profile, teachers can evaluate if that recommended content would help that particular learner or not. For this, teachers need to be thoroughly trained on how to integrate these AI systems in a classroom setting (Karam et al, 2017; Pane et al, 2014)

2. The second important dimension to be considered when selecting a ML model for an AI tool is the explainable and interpretable capabilities of that model. Some ML models are considered more explainable than others (Holzinger, 2018; Bukart and Huber, 2021; Bhatt et al, 2020). In education, interpretable AI is especially challenging because educators are not usually considered very tech savvy. This means the companies developing an AI product to facilitate educators need to go an extra mile to ensure that these explanations are understandable for teachers. Alonso and Casalino (2019) have used the ExpliClas explainable AI tool with the Open University Learning Analytics Dataset (OULAD) to predict students' outcomes and produce graphical and textual explanations of these predictions for different stakeholders involved in the educational process. Putnam and Conati (2019) conducted a survey with university students to evaluate the need for explanations in intelligent tutoring systems and concluded that most university students want to see explanations in such systems. Some ML models like big neural networks have been considered black boxes due to the complexity of calculations that happen inside them but recently a lot of work has been done to make these black box models more transparent (Rudin, 2019; Davidson, 2019; Buhrmester et al, 2019; Matetic, 2019; Liang et al, 2021).
  
3. The third factor that can impact upon the choice of a machine learning model in an AI system is the quantity and quality of data that is available. A lot of data is usually preferred to train ML algorithms, but data collection can be a time consuming and expensive process and shortage of data has been a long-standing problem in education (Dorodchi et al, 2019). From the data processing stage, this issue is usually tackled with synthetic data or class balancing techniques as discussed in the previous section. From the machine learning modelling stage, techniques like *few shot learning* have emerged to build AI systems from small amount of data (Sung et al, 2018; Ravi and Larochelle, 2016; Wang and

Yao, 2019). Wu et al (2021) have shown how *few shot learning* can be used to provide student feedback at scale to around sixteen thousand students. Wu et al (2019) have used zero shot learning for rubric sampling for coding. This research shows the application of cutting-edge ML techniques in education, but the extent to which they help educators needs to be proven. Often these techniques are black boxes that are difficult to interpret by educators which limits the usefulness of these algorithms. As shown above, these techniques have many a times predicted student drop out with a great deal of accuracy. But they do not explain why a particular learner may drop out from a particular course. This information would help the educators to plan the right interventions for such students.

Figure 9 from Baker and Hawn (2021) shows the different stages of an AI development pipeline. They show the different types of biases that can occur in the machine learning development lifecycle according to various researchers (Olteanu et al, 2019; Suresh and Guttag, 2020; Cramer et al, 2019; Paullada et al, 2020; Mitchell et al, 2020; Mehrabi et al, 2019).

In figure 9, the sixth row called 'Model Learning' is where the machine learning model selection, training, testing and evaluation happens. The ML modelling stage also plays a very important role in tackling bias and identifying the limitations of data. The aggregation and evaluation bias shown in the ML modelling stage in figure 9 are enrooted in the data processing stage but are very often identified in the ML modelling stage after analysing AI tool's predictions and decisions. Most of the biases throughout the AI development lifecycle like aggregation and evaluation bias, or deployment and feedback loop bias can be traced back to the basics when the AI tool's development was being strategized and data collection decisions were being taken. The mistakes or loopholes during this step are aggregated as we move ahead with the ML development pipeline. The ML modelling stage can act as a litmus test for the suitability of quality and quantity of data being used to address the problem at hand.

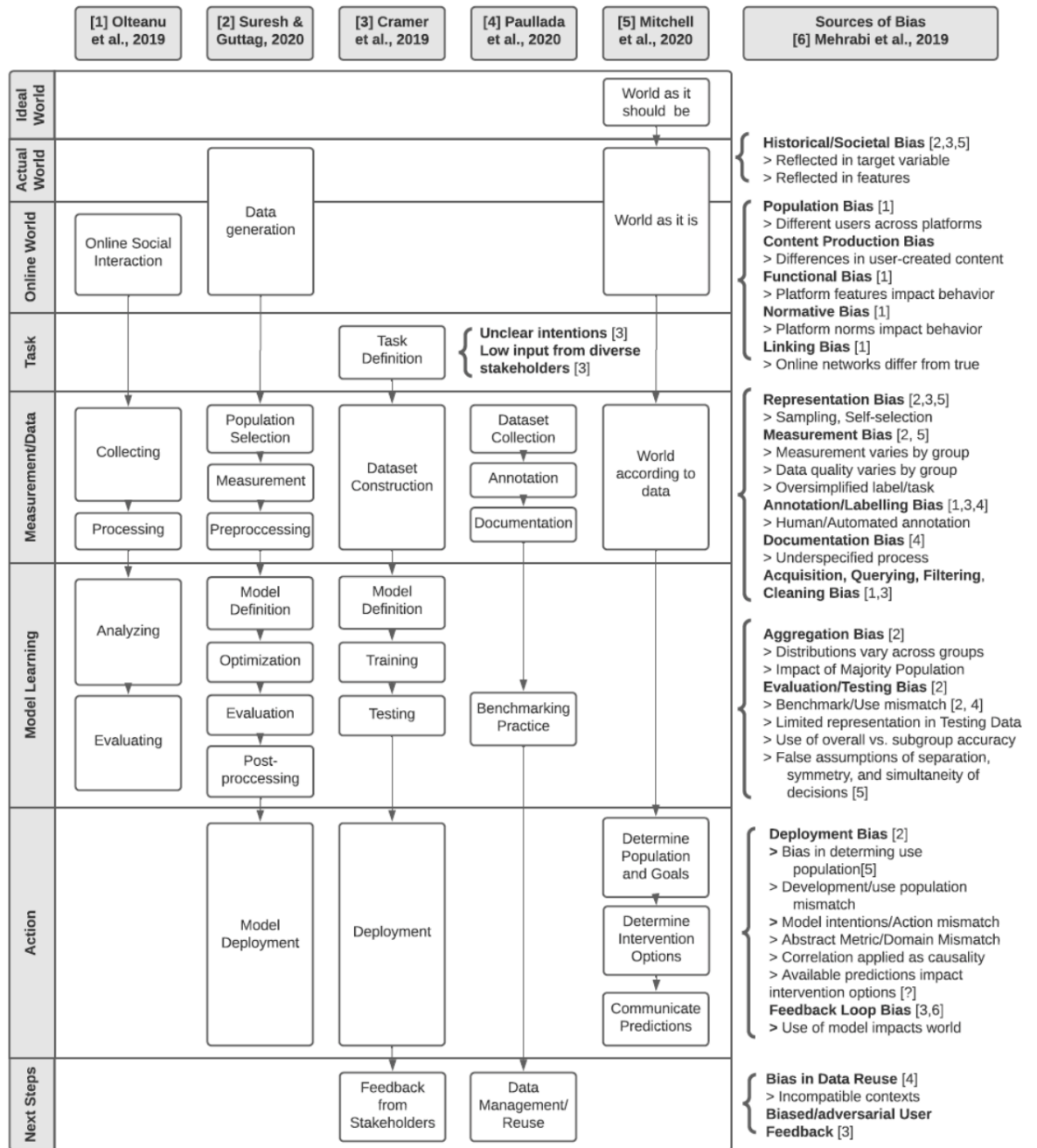


Figure 9: AI Development Pipeline and Sources of Bias (Baker and Hawn, 2021)

The results from ML models on the test dataset can reveal several weaknesses of the AI system. These deficiencies can be limited to the dataset being used or can encompass the model selection or data division sections when the data is divided into training, validation and test sets.

If the ML model performs as expected in the test and validation sets, there are still no guarantees that it will perform as expected in the real world after deployment as discussed in the next section. Deployment bias, feedback loop bias, bias in data reuse and adversarial user feedback illustrate some of the risks when there is a mismatch between the training data of the AI tool and real-world data that tool faces after deployment.

#### **2.8.4 Deployment and Improvements**

Luckin *et al* (2006) have illustrated the importance of human centered design in developing educational systems that are fit for use. They highlight the importance of iterative improvements in building such educational systems. This also holds true for AI systems in education. Work does not stop after an AI system has been deployed in the real-world. It needs constant monitoring and evaluation to ensure that the AI system is performing as expected. As long as there are improvements being made to an AI system in the real-world, transparency considerations should be taken into account. An AI tool might be deployed in the real-world with certain assumptions that may not hold true anymore or may change with time. In such instances, the changes that are made to an AI system should be made transparent for the relevant stakeholders. For example, an AI system used to predict house prices may not take account of new government regulation increasing taxes on house sales resulting in reduced demand and lower prices for houses.

Ideally, the identification of limitations of an AI system should happen earlier than the deployment phase in an AI development lifecycle. But many a times the problems in an AI system are diagnosed only when they are deployed in the real world. This is because there may be differences between the data that an AI system is trained on and the kind of data it faces in the real-world (Xu et al, 2021; Jameel et al, 2020). In some cases, mistakes are incurred in defining the problem and choosing a weak proxy to measure that problem in terms of data. In such contexts, the foundations of an AI system are set on wrong assumptions.

Within AI for education, these limitations were evident in the AI tool used for grading A-level students in United Kingdom during the pandemic (Kippin and Cairney, 2021; Jackson and Panteli, 2021; Smith, 2020; Muers, 2020). It is a typical example of AI gone wrong. From the limitation of the AI-powered grading tool it seemed that it was not tested in real world with a sample of students before being applied nationwide. This AI-powered grading tool was criticized for low accuracy, unfair outcomes and discrimination against students from certain backgrounds and certain schools (Kolkman, 2020).

This tool did not necessarily face different data from the one it was trained on, but it was built on a questionable assumption that mostly did not hold true in the real world. There were some fundamental flaws in the design of the AI tool for making predictions about students' grades. For example, the tool made the A-level results of a student dependent on their school's results in the past three years (Kolkman, 2020). This meant that the creators of the tool assumed that if students from a particular school have not performed very well in the past three years, they are unlikely to perform better this year too. There might be nothing wrong with predicting the future based on the past in some instances. But this assumption in this particular instance enforced an unfair ceiling on students from certain groups by terminating any opportunity for them to perform better than their seniors.

There have been many incidents of AI tools malfunctioning in the real-world. This is because the data they face after deployment is different from the data they were trained on. This phenomenon is known as concept drift (Tsymbal, 2004; Webb et al, 2016; Lu et al, 2018; Xu et al, 2021; Jameel et al, 2020). Even if the bot performs as expected in real-world initially, it may learn adversarial examples or be misused which can produce unexpected results. For example, Microsoft released a chatbot named Tay with the persona of an 18-24 year old American woman in 2016. It could have conversations on Twitter and learn new things based on that. This autonomous ability to learn with no human intervention enabled Tay to learn offensive language, as a

result of which it had to be taken down (Neff and Nagy, 2016; Cetinkaya et al, 2020).

This example shows that the importance of having the right checks and balances in the deployment and improvements stage of an AI development pipeline should never be underestimated. A review of the rise and fall of facial recognition powered AI systems in the real world offers an interesting case study to support this claim. It shows how an optimization of the entire AI development process enabled researchers to achieve state of the art performances in image recognition (Alom et al, 2019) technology in the labs. But, when applied to facial recognition systems in the world beyond the AI labs, it underperformed and seems to have unsolvable bias.

#### **2.8.4.1 Facial Recognition**

Computer vision presents one of the most widely adopted use-cases for AI. It is being used for diagnosing diseases (Mun et al, 2021; Kobayashi, 2019), powering autonomous vehicles (Shreyas et al, 2020; Vinothkhanna, 2020), identifying human emotions (Patel et al, 2020; Meeki et al, 2020), tracing objects in videos (Zhang et al, 2020; Fan et al, 2020) and generating AI-powered content (Kollias et al, 2020; Shoshan et al, 2021). The breakthrough in computer vision got public attention in 2012 from AlexNet (Krizhevsky et al, 2012) when it performed significantly better than other AI systems in the ImageNet Large-Scale Visual Recognition Challenge. Since then, there have been several research breakthroughs and real-world applications of computer vision technology (Iandola et al, 2016; Alom et al, 2018; Hosny et al, 2020). One of its most important applications has been in facial recognition systems around the globe.

Big tech companies like Google and Facebook with dedicated AI labs and almost no budgetary constraints adopted and deployed these systems at scale on their most popular products that are used by hundreds of millions of people around the globe (Buckley and Hunter, 2011). It was also adopted at government level and deployed in schools around the globe (Parsheera, 2019; Zhao, 2021; Liu et al, 2021). But, slowly some of the fundamental flaws

in such AI systems started to emerge across different applications. Google's facial recognition was caught identifying black people as gorillas (Burrell, 2016; Kyriakou et al, 2019; Lohr, 2018). Google apologized and confirmed that it would fix the problem as soon as possible. Three years down the line, it was noted that the company had simply blocked its AI system from identifying gorillas rather than fixing the problem (Vincent, 2018). Facebook had to apologize as its facial recognition system was labelling black men as primates (Demartini et al, 2021). After a few weeks, Facebook announced that it would stop using facial recognition on its platform (Meta, 2021). It is important to note that Facebook's chief scientist is a pioneer of image recognition technology and for decades has played a very important role in taking this field forward. In education, deployment of facial recognition systems inside schools have also faced a lot of criticism (Aljazeera, 2021) over privacy concerns in countries like India. France and Sweden have already banned the adoption of facial recognition in schools (Sunshark, 2021).

Facial recognition presents a typical example of AI systems going wrong after deployment. Hence, it is very important to have checks and balances in place to minimize the adverse effects of AI systems if they go wrong in the real-world. In education, facial recognition has shown promise in authenticating learners in online virtual systems (Valera et al, 2015), e-assessment security for online assessments (Apampa et al, 2010), identifying boredom, confusion or frustration among learners (Dewan et al, 2019) and identifying learners' micro-expressions that may correlate with conceptual learning (Liah et al, 2014).

With the Covid'19 pandemic and shift to online education, all these applications of facial recognition systems in education seem very promising in experimentation phase, but their deployment in real-world settings with students from different ethnicities, color, race, social backgrounds and different mother tongues may pose challenges that have not yet been perceived in the labs.



## 2.9 Gaps in the Literature

In the literature review above, a number of gaps pertaining to transparency in AI development pipelines for ed-tech products were identified.

Firstly, transparency as an important construct of ethical machine learning powered AI has not been widely explored and conceptualized in educational contexts. For example, what does it mean for ed-tech companies developing AI products to make their development process transparent? What kind of transparency is suitable and useful for different stakeholders of AI in education such as educators, ed-tech experts and AI practitioners?

Secondly, there are no theoretical and empirical frameworks for enabling transparency in AI powered products for educational contexts. There are a number of domain agnostic documentation tools covering certain aspects of an AI development process, but there are few frameworks covering the entire AI development pipeline, especially for educational contexts.

Thirdly, there is very limited work on the role of transparency in understanding the gap between digital data collected for machine learning powered AI and human experiences. In addition, there is limited research investigating the concept of transparency with significant input from stakeholders of AI in education to address this gap by highlighting the limitations of collected data in an easily understandable manner.

Fourthly, in educational contexts there have been few evaluations of a transparency framework for machine learning powered AI by different stakeholders like educators, ed-tech experts and AI practitioners. There is almost no transparency framework co-designed with the stakeholders of AI in education.

This thesis aims to address these gaps by presenting a Transparency Index Framework for AI in education that was developed based on a thorough

literature review, a case study of developing and deploying an AI-powered ed-tech tool and interviews with several stakeholders of AI in education.

## **2.10 AI-powered Ed-tech tool**

Ed-tech has been defined in different ways in the literature. Some researchers have limited its definition to tools developed for schools and teachers (Tondeur et al, 2016; Mangal and Mangal, 2019) while others have encompassed tools for various aspects of lifelong learning and professional training as well (Wilson, 1997; King, 2002; Cakiroglu and Atabay, 2022). Januszewski and Molenda (2013) have defined it as *'the study and ethical practice of facilitating learning and improving performance by creating, using and managing appropriate technological processes and resources'*.

The AI-powered ed-tech tool developed in phase two of this research study is conceptualised as an ed-tech tool, because its purpose was to improve the performance of office managers by educating them about the applicants they evaluated for the trading roles. The tool was conceptualised as an ed-tech tool rather than an autonomous AI tool that would make automated predictions or recommendations about hiring or not hiring candidates. Rather, the role of the ed-tech tool was to provide office managers with an extra dimension of (AI-powered) information during the recruitment process about the potential of applicants to perform as traders in future. The office managers were expected to use this information from the ed-tech tool along with their own judgement to reach a final decision. In essence, the tool was designed to “educate” office managers about what good performance looks like in this particular context, rather than making a normative judgment of good performance and executing this instead of office managers.

As an important design decision, the tool was built with an interface that showed the predicted performance measures of candidates so that the office managers could see those and learn from them. It also showed confidence intervals to office managers so they could take a more informed approach to

trust the ed-tech tool's information to be considered, or to ignore it in their own decisions.

The learners of this ed-tech tool were office managers who wanted to learn about the applicants for trading jobs, based on the performance and trading patterns of the company's current traders. The traders previously hired used to make losses for the financial services company during the first few months when they were being trained, and mostly left after the training ended. Based on the initial analysis of the data from the project lead, my research team showed that the financial services company was not always recruiting the most suitable candidates. Hence, office managers wanted to improve the quality of recruitment by learning more about what makes an applicant perform better in light of the company's current traders' performances and their personality traits and behaviours.

The ed-tech tool was powered by AI to predict how an applicant may perform as a trader in the future. It was modelled using the behavioural patterns and trading performance of the company's current traders and predicted the new applicants' potential as traders based on this historic data. After the development of the tool, it was tested with different office managers in enhancing office managers' understanding of successful trader performance as well as looking at the extent to which their recruitment performance improves with the help of the AI-powered ed-tech tool.

## **2.11 Conclusion**

When taking transparency considerations into account while developing and deploying AI systems in the world beyond the lab, it is very important for companies and AI practitioners developing AI systems to highlight the measures they take to evaluate their AI systems after they have been deployed. These measures can be technical, such as how they plan to detect and tackle concept drift in the domain of applications or develop an alarm system to evaluate the AI system after a certain accuracy threshold is reached. If the ed-tech company has done any study or conducted an

experiment in a classroom setting to evaluate or validate their AI-powered tool, it would be useful to also share the context in which this study was conducted so prospective clients of this AI-powered ed-tech product can relate to it.

Such transparency measures would be very effective in diagnosing the limitations of AI tools before they are deployed in production, leading to more robust AI systems. For example, the limitations of facial recognition systems by companies like Google and Facebook would have been identified much earlier if they had analyzed the distribution of different ethnicities and skin colors in the training data of their facial recognition systems. They would have noted significantly lesser data points from non-white skin tones (Lohr et al, 2018). In the next chapter, I cover the importance of such measures and the frameworks that can be used in diagnosing the limitations of AI tools in the data processing stage.

# Chapter 3: Phase 2 - Framework Creation: Data Processing Stage

## 3.1 Introduction

After a literature review, the Transparency Index Framework creation was inspired by my experience of working with a team of researchers to help a financial services ed-tech company utilize AI to learn more about their traders and trading behaviors. This work is presented as a case study to highlight the importance of transparency and application of various ethical AI frameworks on different stages of the AI development process for this AI tool.

The company faced talent-drain as many of their best performing traders left after training. Company executives wanted to learn why traders were leaving after training and how AI could be utilized to predict and understand which traders would leave in the future to enable them to take appropriate steps beforehand. To explore this problem further, it was decided to review in detail the kind of applicants that were recruited by the company, the recruitment process through which they were hired and the kind of training they received after hiring.

In the first step, a Minimum Viable Product (MVP) of an AI-powered ed-tech tool was developed by my team for this company that had the vision to become a leader of AI enabled education technology in the financial sector. The company has 12 offices in 6 different locations and were already becoming increasingly data driven with optimized processes across all offices to collect the most relevant data for the development of AI tools. This work involved a team of researchers working on the new initiative that aimed to harness AI in recruiting and retaining the best traders, training new hires and mentoring the existing traders. All the reports presented in chapters 3, 4 and 5 were prepared by me during the development of this AI-powered ed-tech tool. The financial services ed-tech company's data was used in the development of this MVP ed-tech tool. The tool was to be used by office managers to assist their decision making as they act as 'humans in the loop' when hiring a particular applicant. The role of the ed-tech tool was to enable more informed

decisions by office managers by providing them an extra piece of information based on AI models, without directly recommending or rejecting candidates. The tool was conceptualised as an ed-tech tool to educate office managers with an extra piece of information rather than autonomously taking hiring decisions.

To enhance the understanding of office managers, confidence intervals for predictions (indicating how confident an AI tool is in making a particular prediction) were shown to them as 'humans in the loop' with a more traditional frequentist approach, but such confidence intervals are usually not considered enough for 'humans in the loop' with limited tech expertise to understand and trust the AI systems (Zhang et al, 2020).

As mentioned in section 2.8, there are three overarching questions that need to be considered when planning the development of an AI product. These questions help in determining *how the tool will be used, the types of users who will be using the tool and the way that the tool can impact the lives of stakeholders.*

One of the factors to be considered when strategizing ethical AI development for an ed-tech tool is the tech savviness and value systems of the tool's users (Renz and Vladova, 2021) and those who will be impacted by the tool. Ethical AI development was the top priority when planning the design and development process of the AI tool so the background of people who will be using the tool or impacted by the tool was also taken into account. This helped in determining the resources and time required to make the AI tool's predictions explainable and interpretable for the stakeholders. For the AI-powered tool prepared for the education company in financial services, it was noted that the office managers who will be using the AI tool's predictions are not tech savvy and considerable work needs to be done in presenting the tool's predictions and explanations using lime (Radecic, 2020) to office managers with an easily understandable user interface. Lime is a popular python package used to add explainability to machine learning powered AI products. This experience of identifying the office managers and users of the

AI-powered ed-tech tool as non-technical compared to individuals with a strong technical background like AI practitioners and software engineers, or a bit of technical background like tech enthusiasts, enabled me to categorize different types of users of AI products in terms of their tech savviness. This work was done a priori by another member of my team, but it informed my work on the categorization of stakeholders in the Transparency Index Framework. It showed that these different types of users of AI products have different requirements for transparency. A tool's predictions might be considered fully transparent for an AI practitioner but can be a complete black-box for an end-user with no technical background.

The first step to follow the completion of the data collection process was dimension reduction to identify four different behaviours that encompassed 83.5% of the variance in nineteen different features with respect to trading behaviour. Traders were divided into four clusters based on these four behavioural dimensions:

- Activity of the trader on the trading system,
- Volume and held positions,
- Volatility vs liquid market preferences,
- Diverse vs complex trading.

The performance of traders was then associated with different clusters and each cluster was ranked according to traders' performance (Kent et al, 2021). The clusters acted as 'learning affordances for humans' (Kent et al, 2021) and enhanced the understanding of domain experts like office managers on different traders' behaviours and their association with performance. Based on this work, I further explored which traders from which cluster are more likely to leave the company.

In the next phase, more trading data was received on the actual trades that had been placed by each trader. This included information such as each trade's volume, product name, the type of order, month of the order and whether it was a child order or not etc. Based on this new data, five

behavioural dimensions were identified (including some from previous phase) and traders were divided into clusters according to their trading behaviours.

Movement of traders between clusters was also analysed and significant behaviours that contributed to this movement were identified. T-tests were used to identify the features that played an important role in cluster movement. One of the goals of this wider project was also to develop a tool later for training the traders to help them in becoming better traders and to improve their performance. Change in cluster would be a very important indicator to evaluate the effectiveness of the AI-powered training tool as movement towards a higher performance cluster would mean that the training tool is effective in improving trader's performance.

Based on the initial inferential analysis, I found that the initial cluster (when traders joined the company) of traders is a significant indicator of a trader's performance and may indicate when they will potentially leave the company (leaving is considered a loss for the company because company spends significant amount of resources in training traders, initially traders usually make losses when they join the company and if they leave after a few months when their performance improves, it's considered a loss for the company).

During the recruitment of traders, the financial services company collected data from five different recruitment stages including application forms, questionnaires, math tests, videos submitted by candidates and assessment days that included a face-to-face interview. Additionally, for the AI tool applicants were requested to fill in personality and cognitive questionnaires during their application. Based on this data, each candidate was assigned to one of the four clusters according to the four behavioural dimensions described in this section above.

As a researcher, it was very important for me to get a thorough understanding of the context in which the tool would be deployed. There was an initial study of how traders traded, what constitutes a good trader, their training and the impact of that training. Twenty-eight different features of current traders were



evaluated through personality and cognitive tests. These were then mapped with current traders' trading behaviours and performance. These personality and cognitive tests would also be completed by new applicants, and they would be allocated to one of the clusters accordingly. This work was completed by the project lead before I joined the research team, but I was responsible for documenting the entire process with reflections on what has been learned and what can be improved in future iterations. These reflections and improvements suggested in the documentation became a part of the datasheets shown later in this chapter.

It is also important to note that the company had data being generated from six different locations (UK, Russia, Poland, China, India and Spain) that have different cultures, traditions, beliefs and attitudes towards risk. These differences may be reflected on their trading behaviours. The ed-etch tool was to be deployed across all six locations to empower the office managers across all offices. Some locations like Poland and United Kingdom have a lot more traders than China or India. This means that the ML models were mostly trained on Polish and English traders' data. The location data was not explicitly added as a feature in the model, as this could increase the bias in favour of the traders who exhibit behaviours that are similar to traders in these two locations.

## **3.2 Documenting the Dataset**

The performance of AI-powered products is usually attributed with the quality of data that goes into them (Abedjan, 2022). 'Garbage in, garbage out' is a commonly used term to reflect on the importance of data in determining the performance of AI tools (Kilkenny and Robinson, 2018). Data plays a fundamental role in determining the performance of AI products (Martens, 2018; Sambasivan et al, 2021). Many times, the limitations and biases of an AI system that appear in the machine learning modelling and deployment stages are attributed to the decisions taken and assumptions made in the data processing stage. Hence, transparency of all the decisions taken,

assumptions made and tools used in this stage of the AI development process is very important.

Considering the importance of data in developing impactful AI products, several frameworks and toolkits have been proposed by academics and companies developing AI products to specifically address the ethical concerns of AI in the data collection and processing stage.

The canonical architecture of the AI tool development process in figure 7 shows the sub-components of the AI tool development process: data collection, processing and analytics. They can be the most time consuming and resource intensive processes in the entire AI tool development pipeline (Tae et al, 2019). According to some researchers, this can also be the most challenging part of AI development (Chen et al, 2011; Gitelman et al, 2013).

In education this becomes even more difficult because before starting the data collection process, AI practitioners within educational contexts need to choose proxies for measuring the problem at hand like evaluating students learning outcomes (Skrbinjek and Dermol, 2019; Roberts, 2010) or validating the effects of different pedagogical approaches (Toetenel and Rienties, 2016; Segalas et al, 2010).

Gebru et al (2019) have introduced datasheets for datasets to document the strengths and weaknesses of data used for building AI products. It aims to bridge the gap between data creators and data consumers by documenting the 'motivation, composition, collection process and recommended uses' of the datasets (Gebru et al, 2019). It offers one of the most comprehensive frameworks for documenting different aspects of the data cannon from figure 7 in the AI tool development process. The authors of this research also confirm that the different components of datasheets are not set in stone but they 'expect that datasheets will vary depending on factors such as the domain or existing organizational infrastructure and workflows'. For example, Bender and Friedman (2018), have presented a similar framework to datasheets for Natural Language Processing (NLP) problems.

A datasheet was also prepared for the data used in prototyping and developing the AI-powered ed-tech tool for the financial services company. It covered the documentation of data collection, data processing and data analytics that was followed in developing this tool. It also showed the details of how the dataset was collected, composed, processed and maintained, and the ethical considerations that were taken into account in preparing the dataset.

### **3.2.1 Datasheet for the AI-powered Tool**

The purpose of this document was to serve as a data manual for the training data that was used to train the AI-powered ed-tech tool's models. It provides the details of the data that was already collected, along with its strengths and weaknesses.

It is important to note that I only had 140 traders' data to train the recruitment tool's models. These traders had been with the company for at least nine months. Hence, the data collected is specific to their context and work culture. Data from a large number of traders plus multi-modal data would have helped to improve the ed-tech tool's models and would be acquired in the future, at which point, in those more advanced stages, this specification would be updated.

#### **3.2.1.1 Motivation for Dataset Creation:**

- Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)
  - This dataset was specifically created for this Machine Learning tool, keeping in mind how this data would be used. Extensive literature review was conducted before choosing each category of features for data collection. Table 30 in appendix 1 provides the references to different academic research papers and

articles that provided inspiration to use these features for data collection.

- What (other) tasks could the dataset be used for? Are there obvious tasks for which it should not be used?
  - It should not be used for any domain other than Human Resources or in any sector other than trading. It should specifically be used in this company's context of futures trading. The decision to collect this data was based on a thorough literature review of the relationship between traders' performance and their personality and cognitive traits. This literature review was based on this company's context. As the work culture of companies, even within the same sector like finance can vary significantly, other companies interested in using any models trained on this data, would be advised about contextualizing it to their needs.
  
- Has the dataset been used for any tasks already? If so, where are the results so others can compare?
  - Yes, it has been used during the project for predicting traders' profit and loss, contribution per lot, performance bonus, hard stop counts and clusters (behavioural indicator). Except for clusters, different performance variables are used for routine evaluations of traders. Table 2 illustrates the prediction categories and classes within each category used for prediction.
  
- Who funded the creation of the dataset?
  - Dataset was funded and collected by the financial services company who envisioned to become a leader in ed-tech within their domain. They also funded the translations for different offices across the globe.

### 3.2.1.2 Dataset Composition

- What are the instances? Are there multiple types of instances?
  - It is a survey form that was filled by individual traders and candidates who applied for traders' job opening at the company.
  - It also includes traders' trading data like profit or loss, contribution to the company, hard stop counts, performance bonus and cluster they are in.
  
- What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?
  - Data consists of 28 features and 5 labels that are predicted. In total, there are 140 data instances, each representing a separate trader.
  - Those 28 features were selected from a longer list of features using statistical feature selection methods, applied on English speaking traders' responses. The shorter list was then translated to Polish, Russian and Chinese languages and forwarded to traders with these native languages. Informed consent was taken from traders and the data was pseudonymised so no individual could be identified from the data. The informed consent form used by the financial services company for collecting this data is given in Appendix 4. I accessed this data in a secondary data form.
  - For the training set, I did have access to sensitive variables such as age, gender and office location, for the purpose of validation and identification of biases. The details of these sensitive features is given in the figures 13a-e below, but they were not used for the ML modelling. All the data was pseudonymised before being shared with me so no individual could be identified.

- Most of data is collected in the form of a Likert scale.
  - The distribution of age, gender, location, trading experience and degree category are given in the figures 13a - e below.
- Is everything included or does the data rely on external resources?
    - Data does not rely on external sources, in terms of collection.
    - The content of questionnaires was prepared and validated by academics in the context of different research studies. Table 30 in appendix 1 provides the list of references used. It provides the description of different features that were used for dividing traders into clusters based on their personality. After filling in the consent form, a personality survey was filled by traders to identify certain personality traits that may have correlation with their performance as traders in future.
    - The details of different features used in personality surveys are given in appendix 1.
- Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)
    - We used ten-fold cross validation for optimal results.
- Any other comments?
    - This data consists of data from traders who have been live trading in the company for at least nine months. Based on domain experts' views it was assumed that traders' performance stabilises after nine months, which makes it more likely to be predicted. This makes this data best suited for predictions in this company's context.
    - Different types of survey questions were used to collect the data about the behavioral traits of traders. These questionnaires were designed after a thorough literature review shown in appendix 1. The details of these surveys are given in appendix 2 and some sample questions are given in table 1 and figures 10 and 11 below:

Table 1: Sample survey question

	1 = not at all true	2 = barely true	3 = somewhat true	4 = completely true
When solving my own problems other people's advice can be helpful.				
I try to talk and explain my stress in order to get feedback from my friends.				
Information I get from others has often helped me deal with my problems.				
I can usually identify people who can help me develop my own solutions to problems.				
I ask others what they would do in my situation.				
Talking to others can be really useful because it provides another perspective on the problem.				
Before getting messed up with a problem I'll call a friend to talk about it.				
When I am in trouble I can usually work out something with the help of others.				

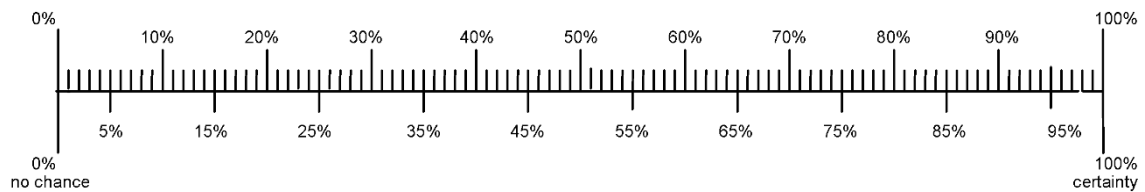


Figure 10 (above): Sample survey question 2: What is the probability that you will keep your permanent address in the same state during the next 5 years?

**Problem**

You have been asked to give a toast at your friend's wedding. You have worked for hours on this one story about you and your friend taking drivers' education, but you still have some work to do on it. Then you realize that you could finish writing the speech faster if you start over and tell the funnier story about the dance lessons you took together.

Would you be more likely to finish the toast about driving or rewrite it to be about dancing?					
1	2	3	4	5	6
Most likely to write about driving	...				Most likely to write about dancing

Figure 11 (above): Sample survey question 3

- Distribution of scores for each feature across the training data were also noted in the datasheet. This distribution gave an overview of the different values for each feature in the raw data that was collected. Figure 12 shows distribution of some sample features that were used in the training data. Appendix 3 shows the distribution all 28 features that were used.

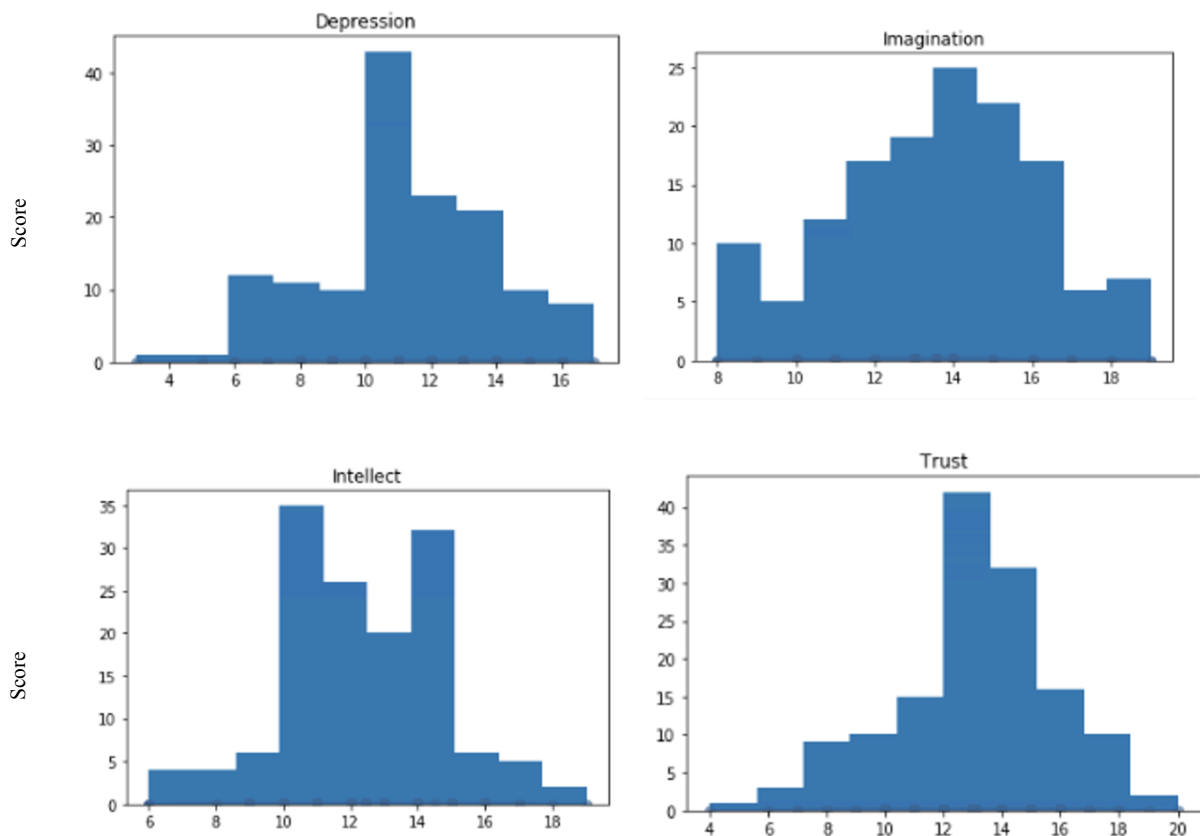


Figure 12: Distribution of scores for each feature used in training data

### 3.2.1.3 Data Collection Process



- How was the data collected?
  - Data was collected by circulating a survey form among the participants (traders) who were also the employees of the company that was collecting the data. Data was pseudonymised and mnemonics were used to identify different instances. Traders provided the data voluntarily, with no external reward and signed the consent form shown in appendix 4.
  
- Who was involved in the data collection process?
  - In the data collection process, the company's new project team, HR team, and office managers across different offices were involved.
  
- Over what timeframe was the data collected? Does the collection timeframe match the creation timeframe?
  - It was collected over a span of three months. Labels data was collected over a span of three years: 2017, 2018 and 2019.
  
- How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?
  - Data was reported by subjects (self-reported) and was not validated/verified. I did account for some outliers and some missing values but there were no consistency checks to double-check if the traders would again fill the questionnaires with similar answers. The surveys themselves were already validated by the time they were used by the company.
  
- Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents). Is this data missing because it was unavailable?

- There is no information missing from the dataset as such. From around 500 traders, only 140 traders' data was used because we did not have more traders' data that have been with the company for more than 9 months.
- Are there any known errors, sources of noise, or redundancies in the data?
  - There are some known deficiencies in the dataset in terms of under representation of certain groups of people. The figures below highlight the distribution of different sensitive variables across the training dataset. The purpose of these figures is to show the potential biases that may exist in models' predictions due to these class distributions. The distribution of dataset with respect to Age, Gender, Location, Degree Category and Trading Experience is given in figures 13a to 13e below

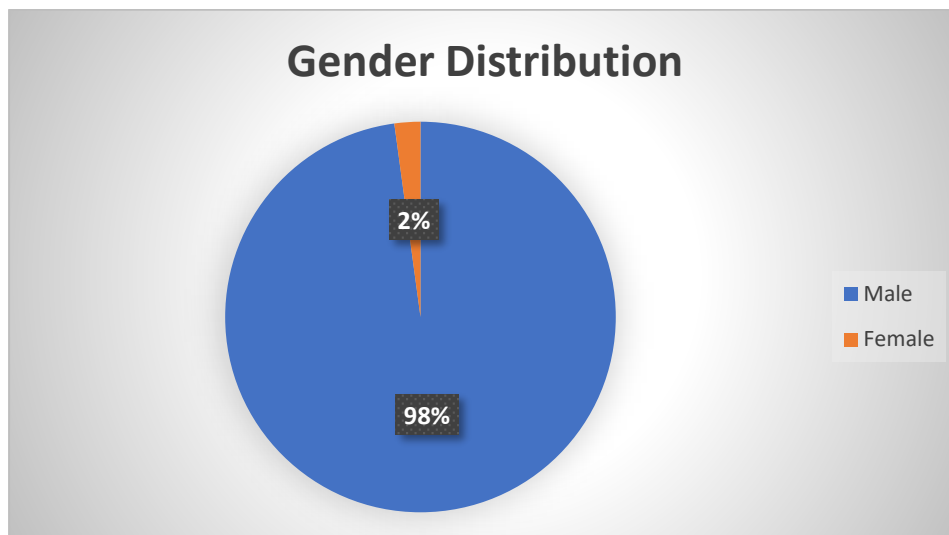


Figure 13a: Gender distribution in the training data of prediction models

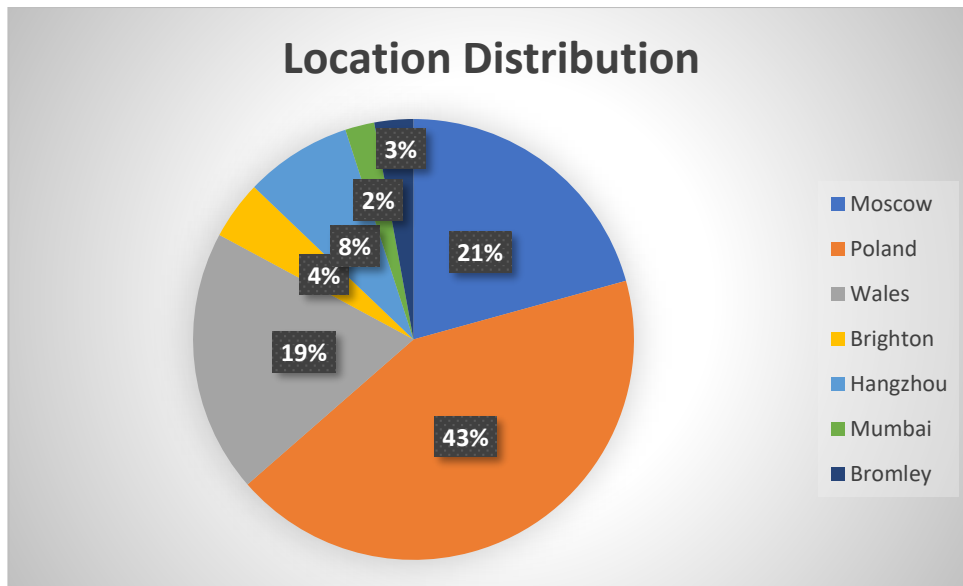


Fig 13b: The company has offices in different cities around the globe. Figure above illustrates each location's distribution in the training data of prediction models

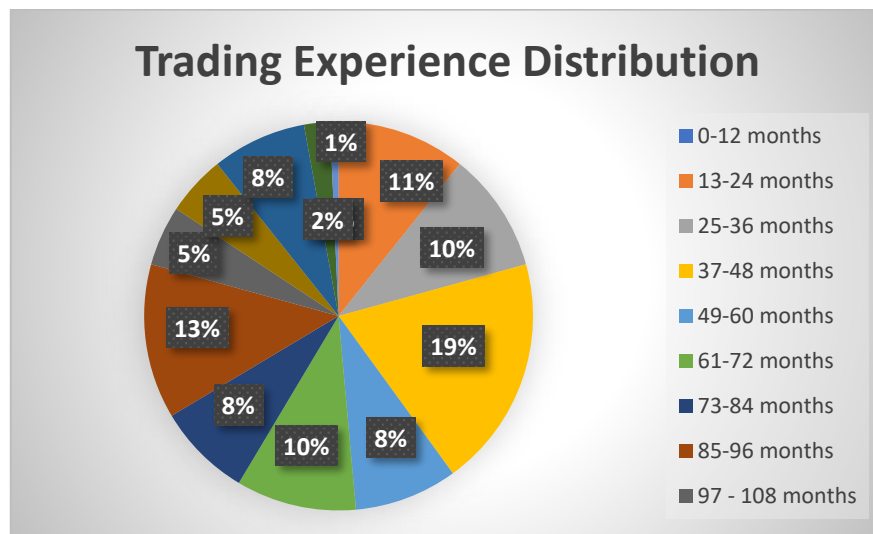


Fig 13c: Distribution of the experience of traders (in months) in the training data

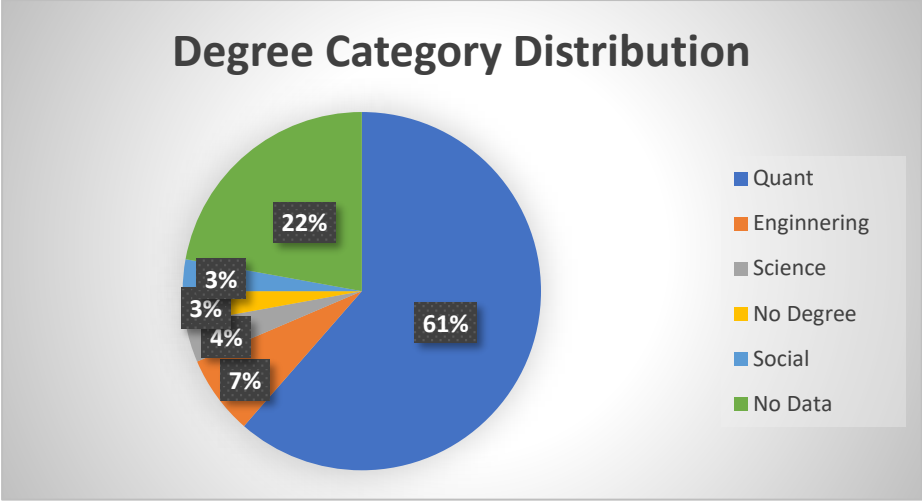


Fig 13d: Distribution of the categories of degrees obtained by traders in the training data

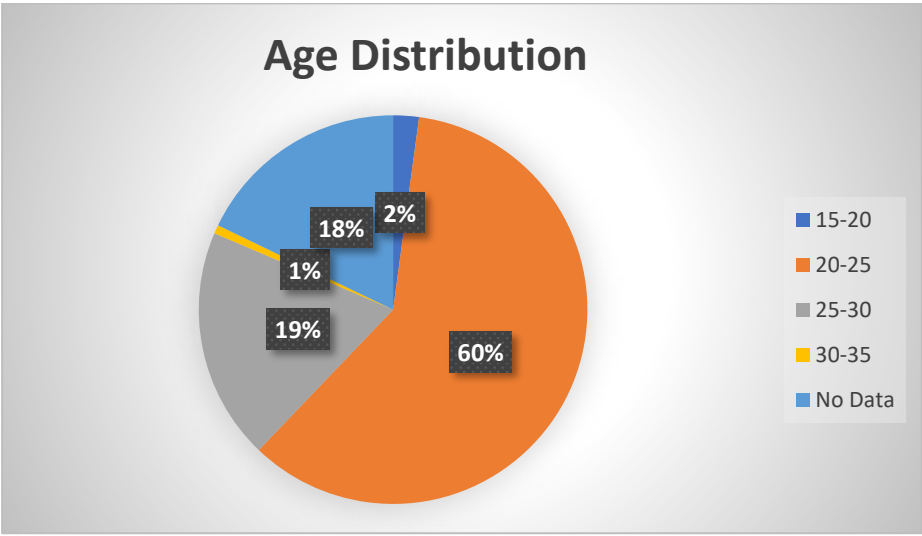


Fig 13e (above): Age distribution of traders in the training data of prediction models

3.2.1.4 Data Pre-processing:

- What pre-processing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)
  - The 28 constructs were collected based on literature review as described in the survey in Appendix 1, then the resulting scores were normalised to a scale of 0 to 1. Some of the models used the normalised data, while for other models non-normalised data was used as it performed better as shown in the model evaluation report in the next chapter. Records in which the responder did not fully filled the survey, or where their mnemonics could not have been linked to performance indicators were omitted from the analysis.
  - To perform classification, labels were divided into different classes based on the domain experts' preferences and number of instances in each class. Table 2 illustrates the classes within each label.

Table 2: Recruitment tool's prediction categories and classes for candidates

Index	Prediction Category	Prediction	Classes		
1	Performance	Profit and Loss (P and L)	Class 1 <= 5000		Class 2 > 5000
2		Contribution Per Lot	Class 1 <= £0.1	Class 2 <=£0.25	Class 3 >£0.25
3		Performance Bonus	Class 1 <=£500		Class 2 >£500
4		Hard Stop Counts	Class 1 = 0		Class 2: >0
5	Behaviour	Clusters	Class 1 = Cluster 1	Class 2 = Cluster 2	Class 3 = Cluster 3 or Cluster 4

- Was the “raw” data saved in addition to the pre-processed/cleaned data? (e.g., to support unanticipated future uses)
  - Yes, the raw data is saved securely under the company’s guidelines.
- Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
  - Yes.

### 3.2.1.5 Dataset Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - The financial services company who owns the data.
- Will the dataset be updated? How often and by whom?
  - The dataset may be updated, as new traders reach the threshold of nine months of live trading.
- If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
  - The data is planned to be updated annually. There is no such process in place yet, as the data is being updated for the first time this year. The company plans to develop such a process.
  - This document (datasheet) would track how often the data is updated and what changes are made periodically to the data.

### 3.2.1.6 Legal & Ethical Considerations:

- If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?
  - Yes.

- If it relates to people, were they told what the dataset would be used for and did they consent? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?
  - Yes, they were told what this data will be used for and they signed the consent form. They could contact the company's HR to get all their data removed.
  
- If it relates to people, could this dataset expose people to harm or legal action?
  - I do not believe so, these are self-reported statements about personality tendencies and their tendency to be subjected to cognitive biases. The data collected by the traders is being used solely for training the prediction models to aid in the hiring process.
  
- If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?
  - Yes, this data has under representation of certain groups based on gender, age group and geographic location, as shown in figures 13a to 13e above. Some of the features are not evenly distributed between different groups, which is reported in this report. These biases are reported thoroughly, and are advised to be taken into consideration when making decisions based on these prediction models.
  
- If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?
  - They are provided guarantees according to General Data Protection Regulation (GDPR).
  
- Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

- Yes, this data complies with GDPR.
- Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)
  - Yes, and the dataset has been pseudonymized with mnemonics before being shared with my research team.
  - The trading data was stored on the company's servers and was accessed by the research team by logging into the remote pc of the company.
- Does the dataset contain information that might be considered inappropriate or offensive?
  - No.

Zliobaite and Custers (2016) have highlighted the importance of collecting sensitive personal data to avoid discrimination in decision making models. This data on sensitive variables helps in determining the context in which the tool should be used and when it should be avoided. Lack of representation from users belonging to certain groups according to gender, race, religion or demographics is one of the major reasons for bias in AI systems (Avellan et al, 2020). If AI practitioners do not take account of these sensitive variables during the data collection and data processing steps, it will be difficult to know what the data lacks, which types of users should avoid using this AI tool or how these limitations of the data can be shown to users through the user interface.

Figures 13a to 13e show the distribution of different sensitive variables like gender, age, location, degree category and the trading experience of the traders in the training data that was used to build the AI-powered recruitment tool for the financial services company. The pie chart in Fig 13a shows the underrepresentation of females in the training data where 98% of data is derived from male traders. This means that the AI tool may not perform very well when presented with female candidates and is potentially biased against them. Hence, it is very important to show this deficiency in the data, and the



AI tool, to the recruitment manager who is using it. Similarly, candidates represented in the training data are mostly between twenty and thirty years old (79%) which means anyone beyond this age group is less likely to receive correct predictions from the AI tool.

The datasheet also took account of the detailed description of each behavioral feature that was used for clustering the traders into different groups. Table 30 in appendix 1 illustrates the research-based approach that was used in identifying the behavioral traits relevant to trading. All 28 features were chosen after a thorough literature review and with consultation from domain experts. This aspect of the datasheet is crucial for making the data processing stage in the AI tool development process transparent and interpretable for third parties, like end-users. It provides the crux of training data, highlighting which features are chosen, and why.

Additionally, the datasheet also gives an overview of the predictions made by the AI tool's machine learning models. It shows the five different types of predictions made by the tool regarding traders' performance and behavior. It shows the classification of different classes for each prediction category and the thresholds for each class within a particular category. Table 2 highlights the prediction categories and classes within each category that were predicted by the recruitment tool.

The datasheet that was prepared for the AI-powered tool provides a thorough overview of the data cannon in [figure 7](#). It shows how the training dataset for the AI tool was prepared, why this particular dataset was used, which performance and behavioral categories were being predicted from this dataset, the distribution of sensitive variables in the dataset, the distribution of different features in the collected data and the potential bias in some features with respect to gender and location. All these factors are crucial for the transparency of AI development pipelines. Datasheet is therefore added as a requirement in the Transparency Index Framework.

The datasheet was added as a requirement for the data processing stage in the Transparency Index Framework because it provides very useful information about how and why a particular dataset was curated. But to make this information useful for end-users like educators, ed-tech experts and AI practitioners working on ed-tech products, some additional measures need to be taken. To take account of the specific needs of end-users in educational contexts, datasheet in the Transparency Index Framework was wrapped around requirements like explicitly indicating the biases that exist in the dataset and the impact they can have on the tool's predictions.

One major addition to the structure of the datasheet in the Transparency Index Framework was the assumptions made while collecting and curating the datasets. These assumptions are particularly important in educational contexts where mostly proxies are used to measure learning outcomes and progress, pedagogical efficacy and drop out predictions. Documenting these assumptions plays a very important role in their identification and validation. Next section discusses the process that was followed in testing one of the most important assumptions based on which the relevant data was collected, and AI tool was built.

### **3.3 Testing the Assumptions**

The design of any AI tool is bound to be influenced by some assumptions. The assumptions on which AI practitioners envision the tool's performance and the assumptions of business leaders as they envision the AI tool's impact and its business value, for example. In education, proxies have been commonly used to measure different learning outcomes and evaluate various pedagogies as discussed in chapter 2. Even simple use cases in education like predicting grades in assessments for learners are based on assumptions that a particular assessment accurately reflects on a student's learning based on his/her context (Dunn et al, 2009; Luckin et al, 2017).

When developing AI-powered products in education, as many assumptions as possible should be tested through experimentation, validated by domain experts like educators and noted and referred to during the deployment and

iterative improvements stage. Testing the assumption can play a very important role in identifying the ‘unknown algorithmic biases’ (Baker and Hawn, 2021) in AI systems or what researchers call ‘unknown unknowns’ in AI systems (Dietterich, 2017; Zhao et al, 2021; Luusua and Yippoli, 2020). Domain experts like educators, ed-tech experts, learners and sometimes parents can be a great resource to validate the assumptions made in the AI development process (Molenaar, I., 2022).

In prototyping the AI-powered ed-tech tool for the financial services company, domain experts (senior traders and office managers with more than ten years of experience) believed that the new traders’ performance has a lot of variability in the first few months but stabilized after nine months. Hence, after nine months it is usually clear to office managers how good a particular recruit was. This was a very important assumption because it could have a huge impact on the data available for training the AI models. It was decided to test this assumption to make sure that the data that was being used to train the models was fit for this purpose. Nine Month Analysis Report below shows how I conducted the exploratory and statistical analysis to test if traders’ performance actually stabilizes after nine months, as assumed by office managers.

### **3.3.1 Nine Month Analysis Report**

#### **3.3.1.1 Monthly Traders’ Performance**

The AI-powered ed-tech tool we developed educated office managers regarding four different indicators of applicants’ future trading performance:

- Profit and Loss after Rebates
- Contribution Per Lot
- Hard Stop Counts
- Performance Bonus

For the predictions of the above indicators, we are using data of traders who have been with the company for at least ten months, based on the assumption that traders’ behaviour and performance is generally being

stabilized after nine months. The purpose of this analysis is to explore the hypothesis that the company's traders performance stabilizes at some point in their learning curve. Currently, based on the data we are using for predictions, we would expect the traders' performance to stabilize after nine months. But, this is an assumption that is tested in this report.

This is a very important assumption to validate for the transparency of the AI development process because it directly affects the quantity and quality of data we can use to train the machine learning models for the AI-powered ed-tech tool. This can in turn impact the performance of the AI tool after deployment.

### 3.3.1.2 Exploratory Analysis

For analysing traders' performance before and after 9 months of them joining the company, I had data of 531 traders in total. This data was from 2017 and 2018 so I shortlisted the traders for whom we had at least three months of data within the first nine months of them joining the company. After this filtering, we had data of 89 traders.

Figures 14a, 14b, 14c and 14d below show the plots for exploratory analysis of different performance indicators. Different coloured lines in the charts below show different traders' data.

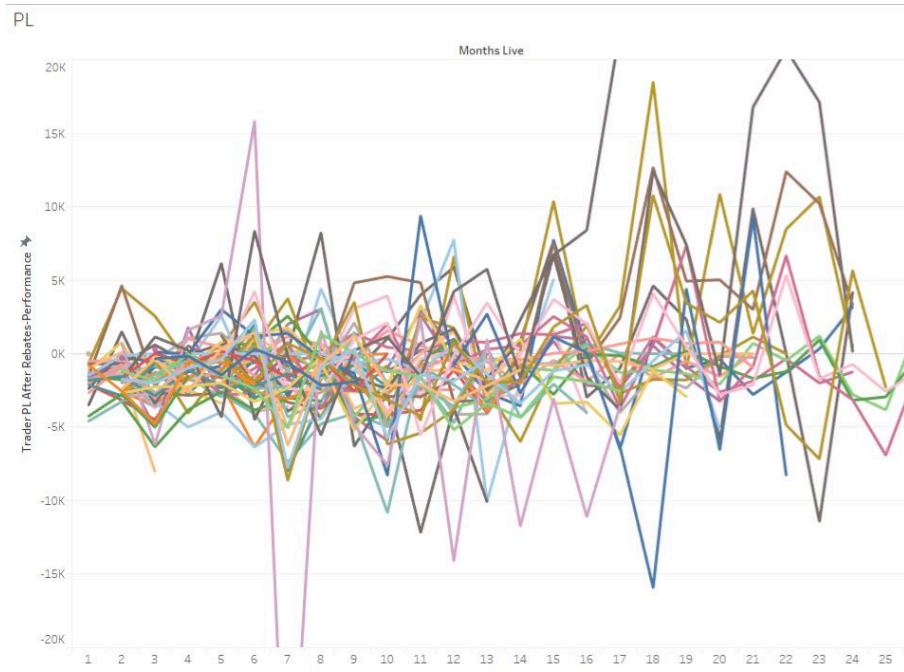
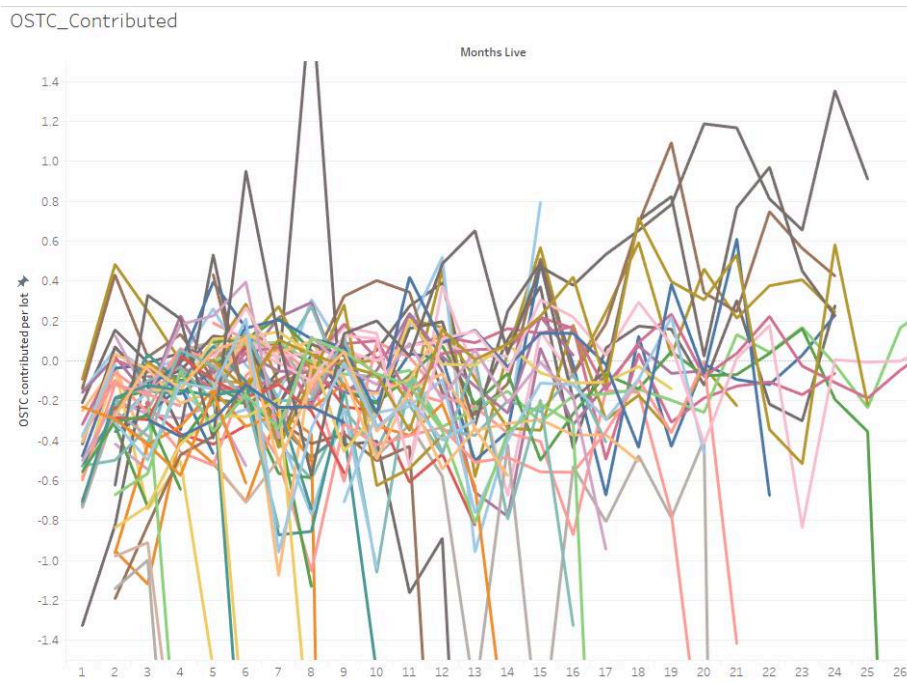
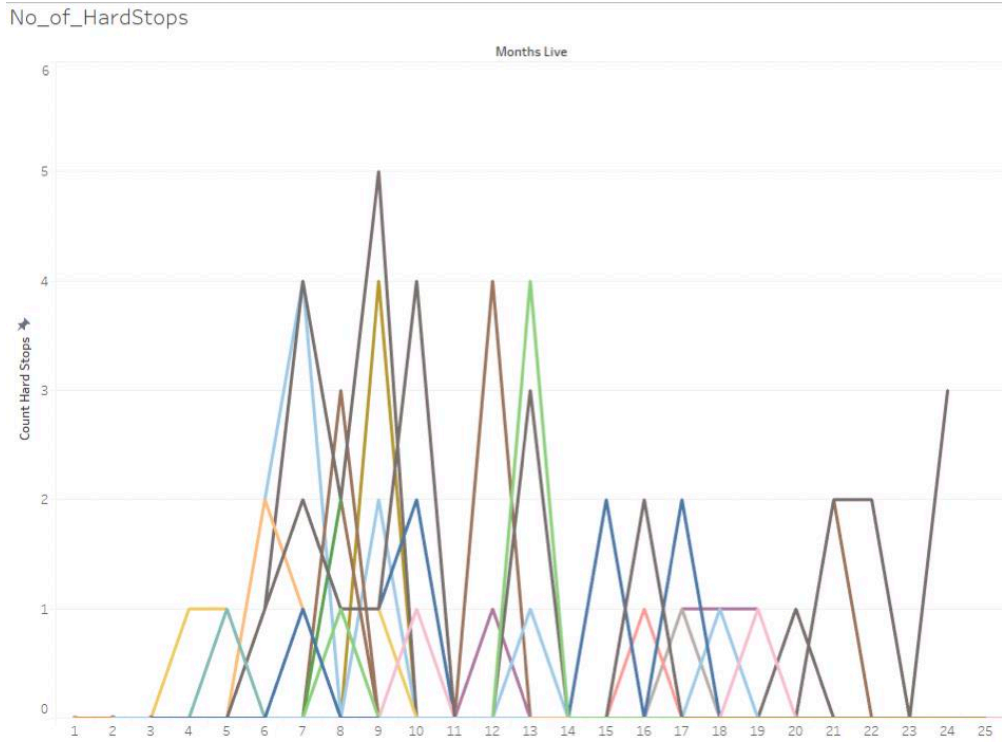


Figure 14a (above): Traders' P & L after rebates



Months Live

Fig 14b (above): Traders' Contribution per Lot



Months Live

Fig 14c (above): Traders' Hard Stop Counts



Fig 14d (above): Traders' Performance Bonus

Table 3: Exploratory Analysis of traders' performance

Performance Indicator	Exploratory Analysis
Profit and Loss	We cannot conclude that traders' profit and loss after rebates becomes relatively stable after nine months of them joining the company (figure 14a above). It actually looks from the chart that the fluctuations are increasing with time.
Contributed Per Lot	We cannot conclude that traders' contribution to the company per lot becomes relatively stable after nine months of them joining the company (figure 14b above).
Hard Stops Count	Variations in traders' Hard Stop counts seems to reduce after 14 months of their joining the company (figure 14c above). In this figure, the number of traders is the same as in other performance indicators, but the values for most months for most traders are zero.
Performance Bonus	We cannot conclude that traders' performance bonus becomes relatively stable after nine months of them joining the company (figure 14d above). In this figure, the number of traders is the same as in other performance indicators, but the values for most months for most traders are zero.

From the above visualizations, based on exploratory analysis we cannot conclude that traders' performance becomes stable after nine months or one year of them joining the company. It seems that variation in traders' performance continues within at least the first two years of their trading careers. This might be because their performance is also dependent on external factors that influence the markets they are trading. Variation in Hard Stop counts seems to reduce after fourteen months, but with limited examples we cannot make this conclusion with certainty.

### 3.3.1.3 Statistical Analysis

From the above exploratory analysis I could not draw any conclusions regarding the variations in traders' performance with certainty. Hence, I resorted to statistical analysis to test the stationarity of time-series data. Our

time series will be considered stationary if its statistical properties do not change with time. In other words, it has constant mean and variance, and covariance is independent of time. If our time series are stationary, they probably will not converge.

Augmented Dickey-Fuller Tests were used to test the stationarity of time-series data for all indicators of traders' performance. The p values and ADF statistic for each indicator are shown in table 4:

Table 4: Augmented Dickey-Fuller Test results for performance indicators

Performance Indicator	Augmented Dickey-Fuller Test Results Analysis
Profit and Loss	P < 0.01, (5.18 e <sup>-13</sup> ), ADF Statistics: -8.258
Contributed Per Lot	P < 0.01, (0.00), ADF Statistics: -27.154
Hard Stops Count	P < 0.01, (1.20 e <sup>-24</sup> ), ADF Statistics: -13.181
Bonus	P < 0.01, (1.43 e <sup>-8</sup> ), ADF Statistics: -6.46

From the p values and ADF statistic shown in table 4, we can reject the null hypothesis and conclude that the time series data is stationary and non-stochastic. This means the performance indicators' data can be modelled and is eligible for predictions, but on the other hand, it would be less useful for convergence. The data does not vary randomly with time and there seems to be no time-dependent structure in our data which can enable us to make useful predictions about traders' performance in future.

#### 3.3.1.4 Auto Correlation Plots

The auto correlation plots were used to confirm that the data was not random and could be utilized for predictions. Auto correlation is the similarity between different observations as a function of time lag between them. It is a representation of the degree of similarity between a given time series and a lagged/later version of itself. We plotted autocorrelation coefficients with different lag values for each performance indicator and analysed them.

Ideally, we want a high autocorrelation between adjacent and near adjacent observations. For our time series data to be suitable for predictions, we



expect higher autocorrelation coefficient values for smaller lags (in months) and lower autocorrelation coefficient values for larger lags. It is important to note that the autocorrelation plots tell us if the data can be used for time series predictions of traders' performance, although for our AI tool, I am not making time series predictions. It needs to be documented and taken into account to ensure transparency for the team working on this tool.

The charts in figure 15a, 15b and 15c below show the correlograms for different performance indicators. These charts illustrate four sample traders' data for each performance indicator, but to draw conclusions in table 5, I evaluated all traders' figures for each performance indicator. In these figures below, x-axis shows different lag lengths in trading months and y-axis plots the auto correlation coefficients.

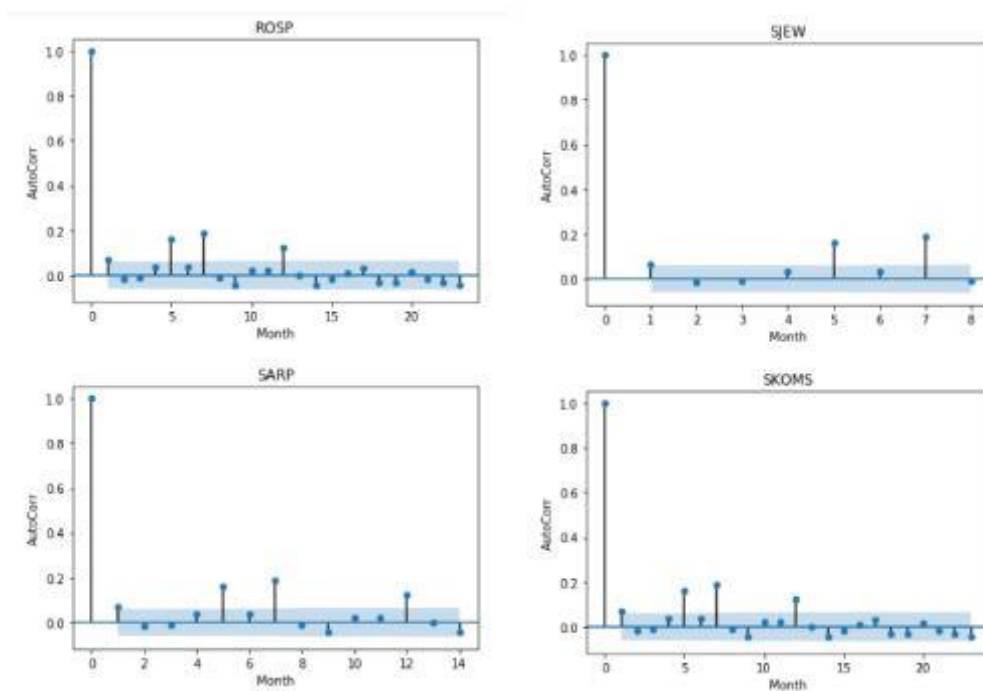


Figure 15a (above): Auto Correlation plots Contribution Per Lot

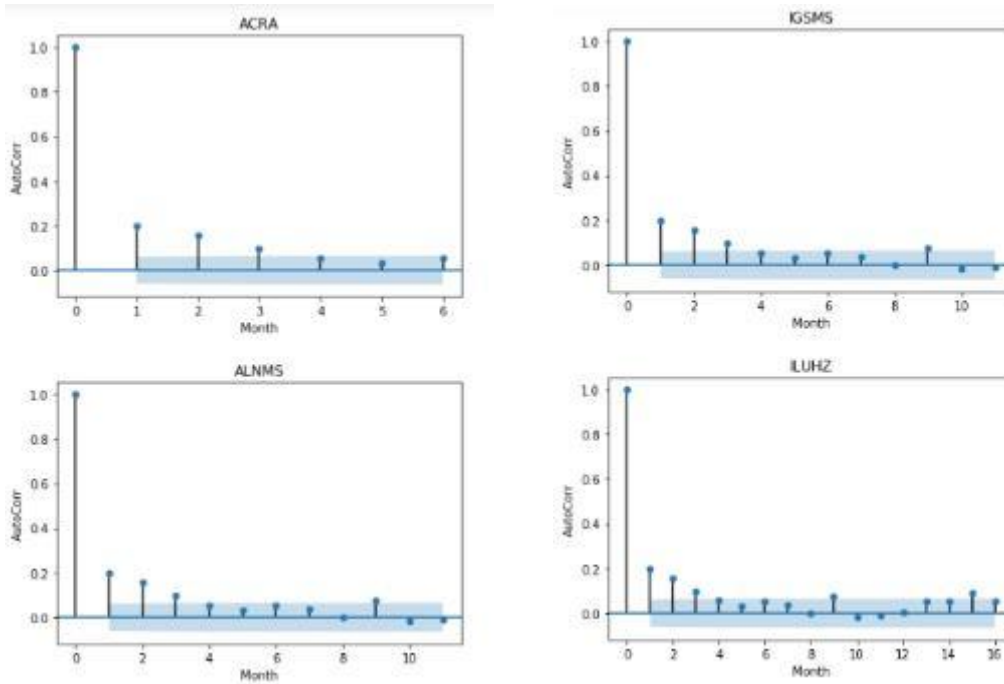


Fig 15b (above): Auto Correlation plots for Hard Stop Counts

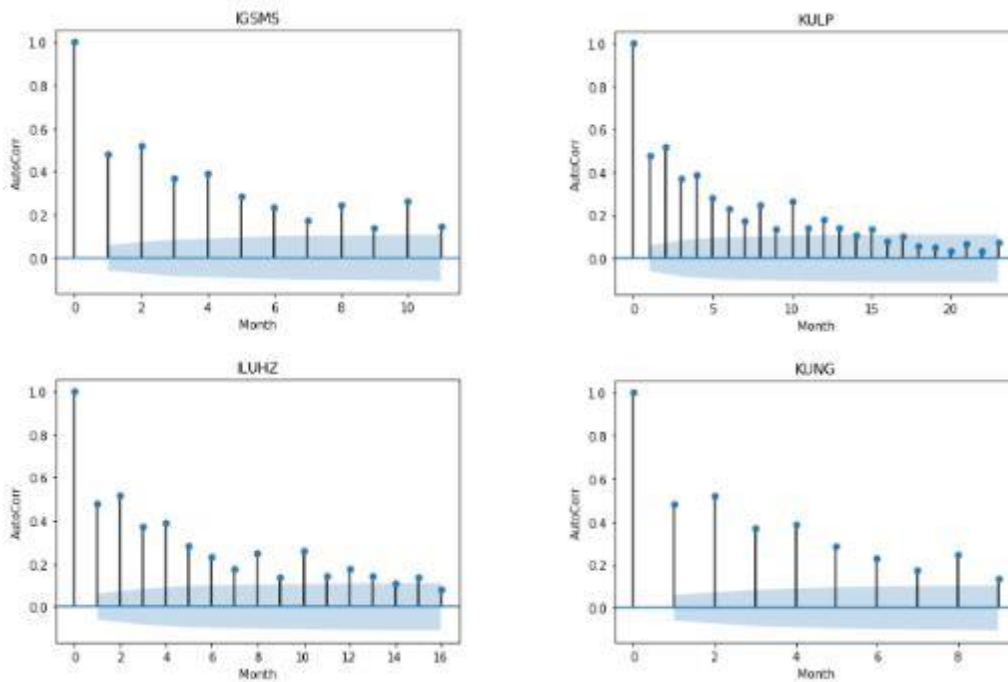


Fig 15c (above): Auto Correlation plots for Profit and Loss after rebates

Table 5 (below): Auto Correlation Plot Analysis

Performance Indicator	Auto correlation Plots Analysis
Profit and Loss	For P & L I observed higher autocorrelation coefficients for smaller lags and vice versa. It shows that auto correlation is high between

	adjacent and near-adjacent values. This means P & L time series data in fig 15c above will be effective for predicting candidates' P & L.
Contributed Per Lot	For Contribution per lot in fig 15a, I could not observe similar patterns as P & L among autocorrelation coefficients and lags. In some months, I saw a high auto correlation value but that can be attributed to randomness. This means Contribution per lot data may not be very useful for making predictions out of time series analysis.
Hard Stops Count	For hard stop counts, I observed higher autocorrelation coefficients for smaller lags and vice versa in fig 15b above. This means hard stop counts time series data will be effective for predicting candidates hard stop counts. Though it is not as effective as P and L data because auto correlation between adjacent observations is less high.
Performance Bonus	For Performance Bonus, a vast majority of traders had values equal to 0 for almost every month, which makes the variance in data negligible. With no variance in data, autocorrelation coefficients cannot be calculated. Hence, I did not have autocorrelation plots for performance bonus.

### 3.3.1.5 Consecutive Month Differences

I also plotted the differences between consecutive months for each trader's performance to analyse if the differences between traders' performance are increasing or decreasing with time. For traders' performance to be stable after a certain month, I would expect the plot values to rotate near zero after some point. These figures 16a, 16b, 16c and 16d below illustrate four sample traders' data for each performance indicator, but to draw conclusions in table 6, I evaluated all traders' figures below for each performance indicator. In these figures below, x-axis shows the month values and y-axis plots the difference between consecutive months in dollars.

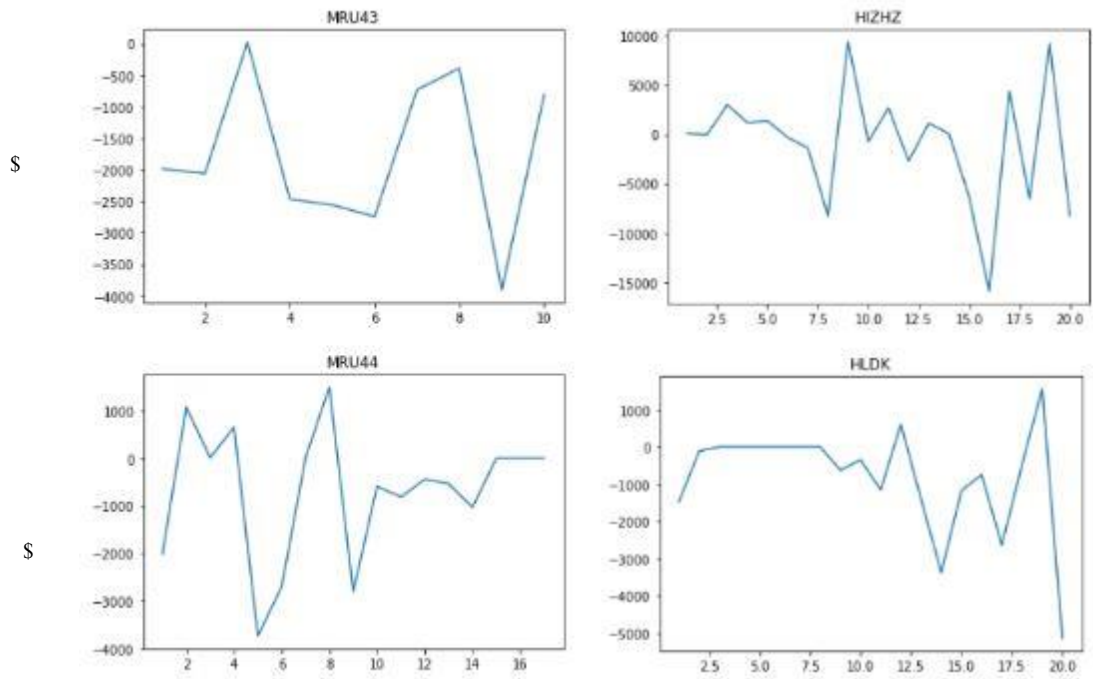


Figure 16a (above): Profit and Loss monthly difference (\$ on y-axis)

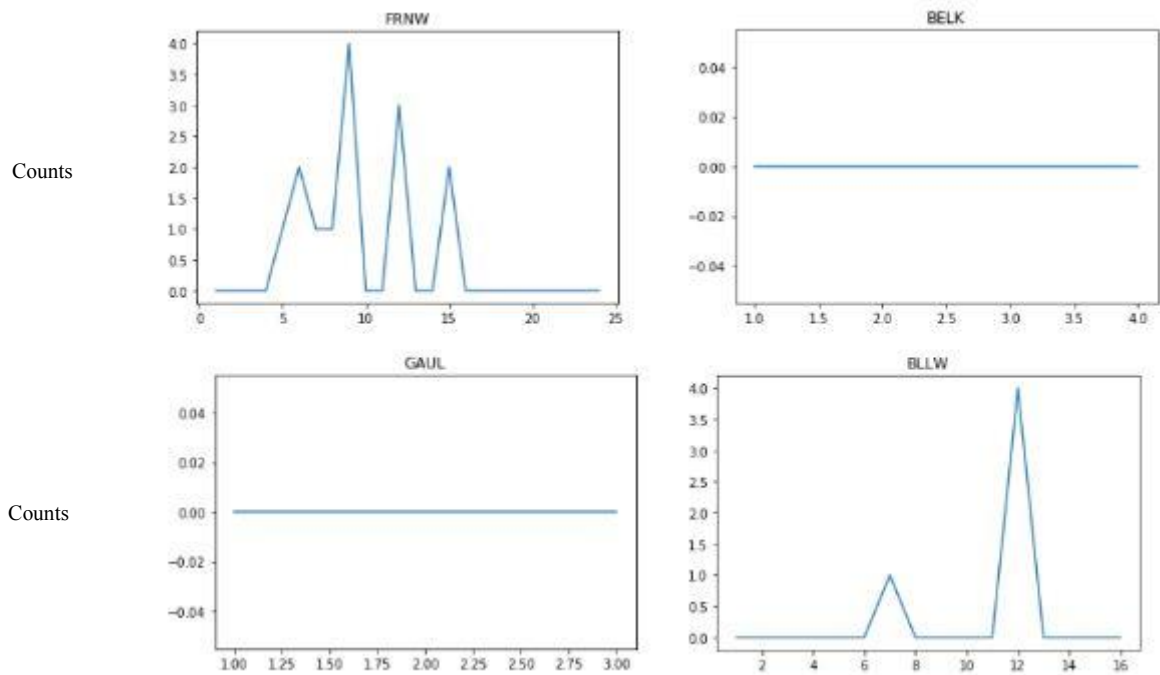


Fig 16b (above): Hard Stop Count monthly differences (counts on y-axis)

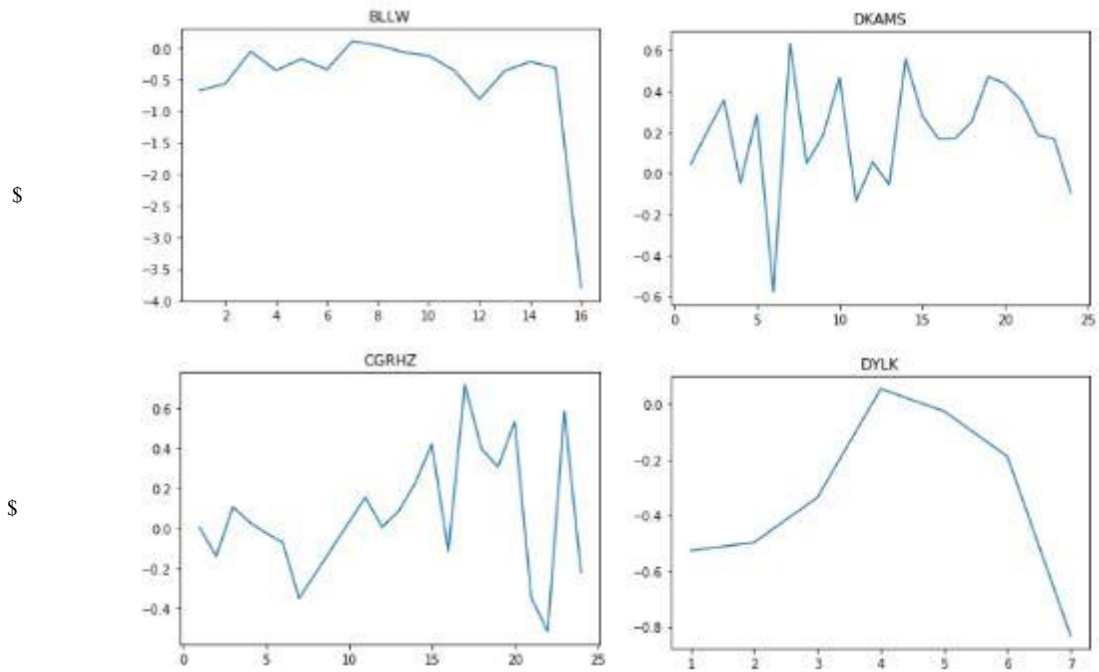


Fig 16c (above): Contribution per lot monthly differences (\$ on y-axis)

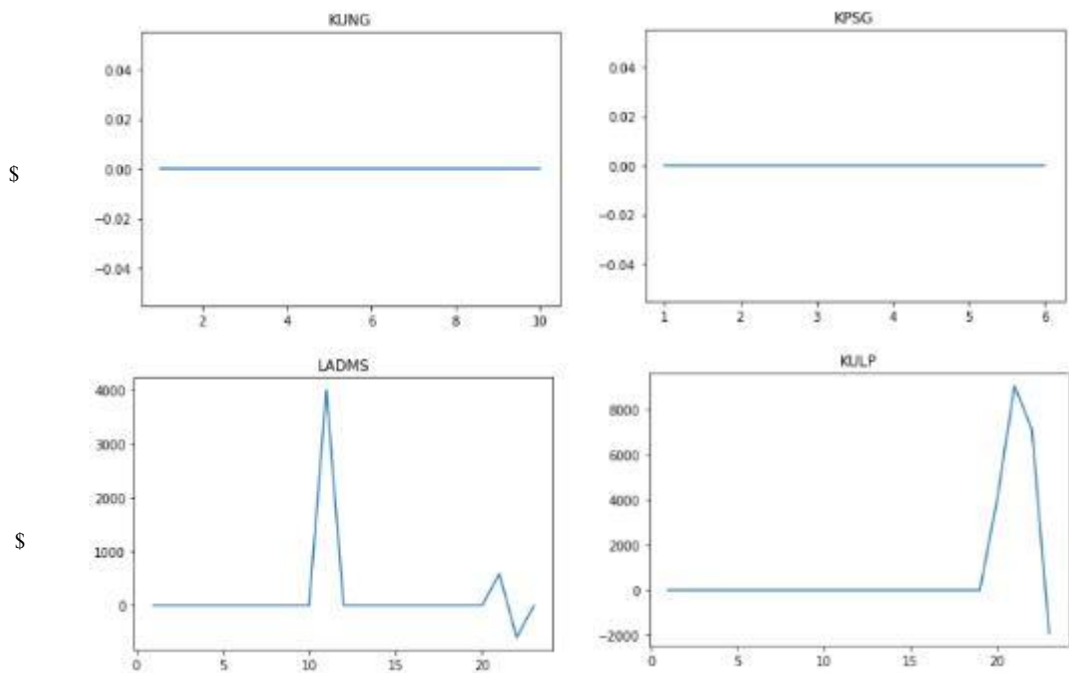


Fig 16d (above): Performance Bonus Monthly Differences (\$ on y-axis)

Table 6: Consecutive Month Differences Analysis for Performance Indicators

Performance Indicator	Monthly Differences Analysis
-----------------------	------------------------------

Profit and Loss	From the monthly differences analysis for traders' P & L, I cannot identify a month after which traders' P & L performance stabilizes. However, at least for some of the traders, it looks like they become <b>less</b> stabilized with time.
Contribution Per Lot	From the monthly differences analysis for traders' Contribution Per Lot, I cannot identify a month after which traders' Contribution Per Lot performance stabilizes. However, at least for some of the traders, it looks like they become <b>less</b> stabilized with time.
Hard Stops Count	From the monthly differences analysis for traders' hard stop counts, I cannot identify a month after which traders' hard stop counts performance stabilizes. In general, quite a few traders did not have any breaches so no difference at all.
Performance Bonus	From the monthly differences analysis for traders' Performance Bonus, I cannot identify a month after which traders' Performance Bonus stabilizes. In general, quite a few traders did not have any bonus so no difference at all.

From the monthly differences analysis, I cannot identify any month for any of the four performance indicators after which we can expect the traders' performance to stabilize.

### 3.3.1.6 Summary Statistics

From the above analysis, we have established that the time series data is stationary. But I have not identified a month after which traders' performance stabilizes. For each performance indicator, ideally I need to find a month after which we expect the variation in traders' performance to reduce. To identify this particular month, among the (mostly) 24 months data that was available, I compared the variation in traders' performance before and after any particular month. Levene Tests were used to compare the variance in traders' performance before and after each month.

Table 7: Levene Test results for performance indicators

Performance Indicator	Levene Test Results Analysis
Profit and Loss	From the exploratory analysis of the results of Levene tests for each month and each trader, variance was significantly different before and after 11 <sup>th</sup> month for most traders, using the significance level of 0.05. The variance in profit and loss increases after 11 months of their joining the company.
Contributed Per Lot	From the exploratory analysis of the results of Levene tests for each month and each trader, variance was significantly different before and after 9 <sup>th</sup> and 10 <sup>th</sup> month for most traders, using the significance level of 0.05. The variance in contribution per lot increases after 9 months of their joining the company.
Hard Stops Count	For Hard Stops Counts, from exploratory analysis of the results of Levene tests for each month and each trader, I could not reject the null hypothesis with a significance level of 0.05 or lower. Hence, no month could be identified after which we expect the hard stop count for traders to stabilize.
Performance Bonus	For Performance Bonus, from exploratory analysis of the results of Levene tests for each month and each trader, I could not reject the null hypothesis with a significance level of 0.05% or lower. Hence, no month could be identified after which we expect the hard stop count for traders to stabilize.

Based on Levene Tests, for two performance indicators (profit and loss after rebates, contribution per lot) I have identified the months after which I can compare the AI tool's predictions with the applicant's actual performance. This is a very important insight to take into account when evaluating the performance of the AI tool and hence needs to be documented in a transparent manner for re-consideration after the AI tool has been deployed.

### 3.4 Conclusion

In this report, firstly exploratory analysis was conducted by plotting and analysing traders' performance. The Augmented Dickey-Fuller test and auto

correlation plots (correlograms) were used to confirm if the data is stationary and suitable for time series and other predictions. Then monthly differences plots and Levene tests were used to identify a particular month after which we can expect traders' performance to stabilize. This would enable me to compare the AI tool's predictions with applicant's performance after they've spent a certain number of months at the company.

From the above exploratory and statistical analysis of 89 traders' data, it was concluded that the time series data for the four performance indicators (profit and loss after rebates, contribution per lot, hard stop counts, performance bonus) is not random or stochastic. This data is suitable for predicting new applicants' trading performance. From the autocorrelation plots, it was concluded that time series predictions can be made for performance indicators like profit and loss and hard stop counts. However, the sweet spot in which performance of traders is being stabilized could not be found. Even more so, it seemed that in two of the performance indicators – there is a point in which data becomes even more variant than it was before. This is contrary to some of domain experts' opinions. These results highlight the importance of testing assumptions when developing AI-powered ed-tech products. Hence, documenting the assumptions and any measures taken to test them form an integral part of making AI development process transparent.

The nine-month analysis report showed that the performance of traders does seem to be very dynamic. From this it can be hypothesized that in general, traders are not very adaptive to market changes, which causes their performance to be very variant. It refutes the assumption made by some domain experts that traders performance stabilizes after nine months. The fact that a stabilization point could not be found does not mean though that all trading months (from the moment traders are going live) should be used for prediction. Although it has been established that trading performance is not being stabilized, we can still maintain the hypothesis that trading in the first months is less representative of their long-lasting trading performance.

This nine-month analysis report was prepared to test only one assumption regarding the stabilisation of traders' data after nine months. From the results



of this report, this assumption did not seem to hold true. The process of exploratory and statistical analysis that was followed to test this assumption was added as a part of the TIF to test various assumptions when developing AI-powered ed-tech products.

The data sheet and nine-month analysis report prepared for the ed-tech company in financial services highlights the importance of testing the assumptions when developing AI-powered ed-tech products. It is possible for the data to differ from the viewpoints of domain experts. This is where AI can play a very important role in teaching domain experts about their area of expertise. To illustrate the importance of AI and its development process as 'learning affordances for humans' (Kent et al, 2021), this research was presented in International Conference on Artificial Intelligence in Education. This research specifically highlighted the importance of transparency in enhancing the understanding of stakeholders regarding their domain.

Datasheet that I prepared during this case study highlighted how transparency for the data processing stage of the AI development process could be achieved. Various sections within the datasheet like dataset creation and collection process or ethical and legal considerations taken into account can have a major impact on the performance of AI-powered ed-tech products. Subsequently, nine-month analysis report showed how and why assumptions made during the data collection process should be documented, and if possible, tested.

After the data has been collected, cleaned and processed, it goes into machine learning models to achieve a particular objective like making a decision, predicting an outcome or classifying into groups. The next chapter discusses how transparency considerations can be taken into account during this stage of the AI development process.

# Chapter 4: Phase 2 - Framework Creation: Machine Learning Modelling Stage

## 4.1 Introduction

After the data is cleaned, processed and analysed, it goes into a machine learning model which then makes predictions based on the input data. Thomas et al (2019) have proposed a machine learning framework for designing and developing machine learning models where the burden of addressing any bias in the machine learning pipeline is on the AI practitioners. The purpose of such tools is to open the black box of AI and make algorithms more transparent for researchers and developers. Koshiyama et al (2021) presented four different verticals of algorithmic auditing that come under the umbrella of Trustworthy AI (Brundage et al, 2020). They include performance and robustness, bias and discrimination, interpretability and explainability and algorithmic privacy.

There are many different machine learning techniques and algorithms that can be used for building AI systems as shown in figure 17. For example, we need to choose between regression techniques (Cui and Gong, 2018) for predicting exact values, or classification techniques (Kotsiantis et al, 2007) to predict classes. The data is then prepared according to the technique and ML model that is chosen.

Choosing the appropriate machine learning algorithm for a particular problem can have a huge impact on the performance of the tool. Every machine learning algorithm has different strengths and weaknesses, and three considerations need to be taken into account when choosing a particular algorithm:

- 1) The accuracy metrics of the model with collected data. For example, recall, precision or the F1 score for classification problems and the mean squared error, the mean absolute error or the R-squared error for regression problems (Gunawardana and Shani, 2009).
- 2) Whether we want the machine learning model to be explainable or not. Some models are considered black box, for example, neural networks

(Castelvecchi, 2016). Others are considered more suitable for explainability (Petkovic et al, 2018, Singla and Biswas, 2021) such as Random Forest (Petkovic et al, 2018).

- 3) The kind of data being used. For example, the size and quality of the dataset. Some models perform at their best with datasets that have more features (Domingos, 2012, Aggarwal, 2018) while others perform better with large amounts of data.

In the development of AI-powered tool built for the ed-tech company in financial services, the random forest algorithm was chosen for predictions, after a thorough review process. I prepared the 'Models Evaluation Report' to evaluate different machine learning algorithms on the dataset that was curated for the tool. Three machine learning models, neural networks, support vector machines and random forest were evaluated based on the three considerations mentioned above. All three models were trained and tested on the training data for the five prediction categories:

- performance bonus,
- profit and loss,
- contribution per lot,
- hard stops and
- behavioral clusters.

The model evaluation report that was prepared for the AI-powered tool enables systematic selection of the appropriate machine learning models to solve a particular real-world problem. This selection process can play a very important role regarding the performance of the tool, its transparency, and hence its impact on the real-world. Therefore, the model evaluation report can be considered an important component of the design framework for the transparency of AI-powered ed-tech tools. A sample model evaluation report that was prepared for the AI-powered ed-tech tool for the financial services company is presented below.

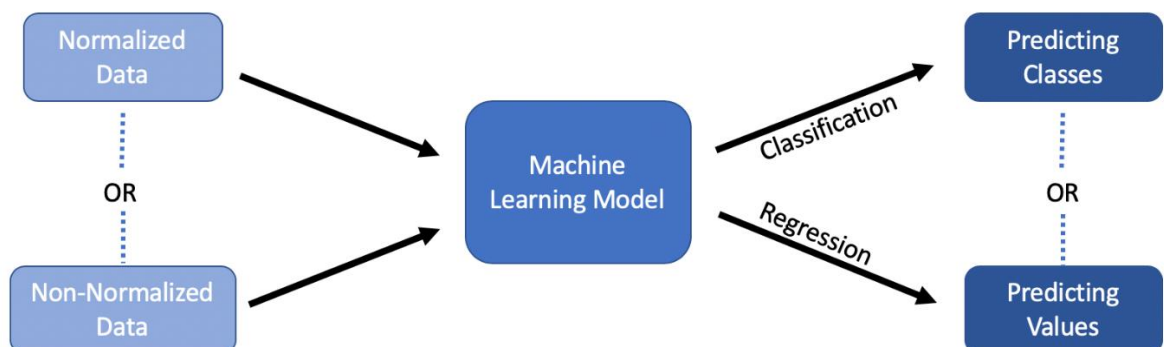
## 4.2 Models Evaluation Report

### 4.2.1 Introduction

The goal of this report is to evaluate and compare the accuracy of machine learning models used for predictions in the AI-powered tool. These accuracy measures might be affected by the kind of data we use, the type of machine learning model we use and types of predictions we make.

Five different models were used to predict the following different categories of traders' performance and behavior:

- **P&L after Costs and Rebates** (monthly average): class 1  $\leq$  £5000; class 2:  $>$ £5000 - (two classes)
- **Contribution per lot** (monthly average): class 1  $\leq$  £0.1; class 2  $\leq$ £0.25; class3  $>$ £0.25 – (three classes)
- **Performance bonus** (monthly average): class 1  $\leq$ £500; class 2  $>$ £500 – (two classes)
- **Hard Stop breaches** (monthly average): class 1 = 0; class 2:  $>$ 0 – (two classes)
- **Clusters**: class 1 = cluster 1; class 2 = cluster 2; class 3 = cluster 3 or cluster 4 - (Three classes, but class 3 predicts the behaviour could be in cluster 3 or 4)



**Possible Paths:**

- Normalized Data and Predicting Classes
- Normalized Data and Predicting Values
- Non-Normalized Data and Predicting Classes
- Non-Normalized Data and Predicting Values

**Possible ML Models:**

- Random Forest
- Neural Networks
- Support Vector Machines

**Possible ML Techniques:**

- Classification
- Regression

Figure 17: Different paths of a Machine Learning Development Pipeline

Figure 17 above shows three parts of the predictions' engine: data processing stage, machine learning modelling stage and predictions stage. In the data processing stage, I had two options to choose from:

- Normalized data: values are restricted between 0 and 1 to reduce the effect of extremely large or extremely low values on predictions as some models are sensitive to extreme values.
- Non-Normalized data: true values are used for predictions as they are recorded, without any restrictions.

### 4.2.2 Machine Learning Models

In the machine learning modeling stage, I tested three different machine learning algorithms for predictions: random forest, neural networks and support vector machines as shown in figures 18a, 18b and 18c. These algorithms were used as they are specifically relevant to the amount and type of data collected.

### Random Forest Machine Learning Algorithm

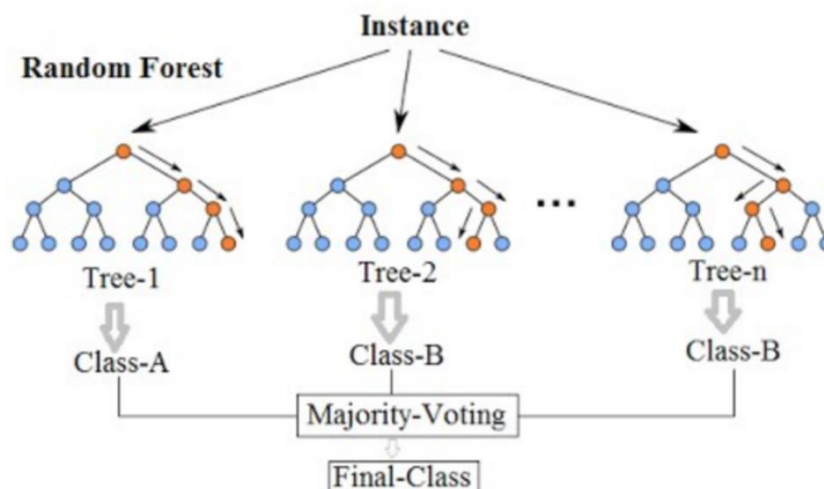


Figure 18a (above)<sup>13</sup> : Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

### Neural Networks Machine Learning Algorithm

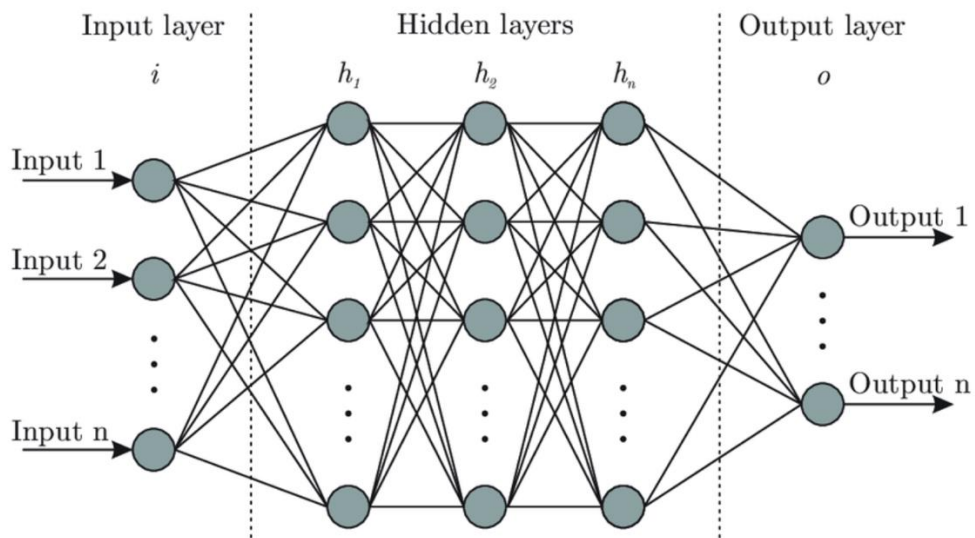


Fig 18b (above)<sup>14</sup>: Also known as Deep Learning, neural networks take features as input, multiply them with weights in hidden layers to identify patterns in the data, and then make predictions based on these patterns.

<sup>13</sup> <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

<sup>14</sup> [https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o\\_fig1\\_321259051](https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051)

## Support Vector Machines Machine Learning Algorithm

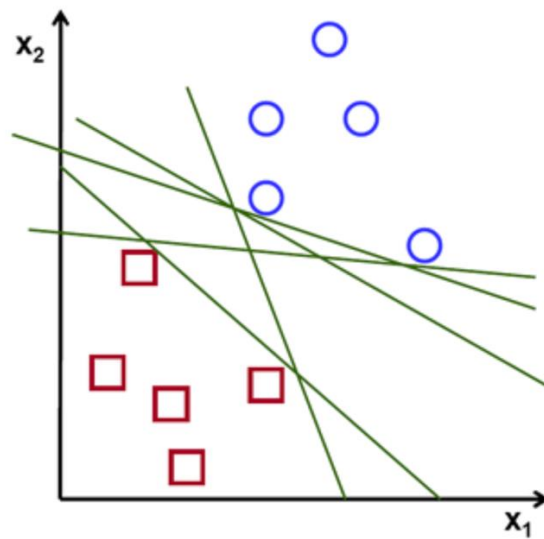


Fig 18c (above)<sup>15</sup>: Support Vector Machines (SVMs) find a hyperplane in an  $n$ -dimensional space ( $n$  - the number of features) that distinctly classifies the data points. This figure shows classification (green lines) based on two features  $x_1$  (red squares) and  $x_2$  (blue circles), in our training data. I had 28 features from the questionnaires data

In the predictions stage, there are two options:

- Regression: predict the exact value of a particular category
- Classification: predict the class or a range of values of a particular data point in that category

All the machine learning algorithms mentioned above need to be optimized by choosing the right set of hyper parameters. These hyperparameters are set before the computations in the machine learning models begin and depend on the kind of problem that is being solved. For example, in random forest one of the hyper parameters is the number of trees, in neural networks one of the hyper parameters is the activation function for each neuron or the number of

---

<sup>15</sup> <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

neurons in hidden layers and in support vector machines one of the hyper parameters is the type of penalty used (L1 or L2).

In the context of the ed-tech tool for recruiting traders, five different models were used to make predictions about five different categories, the hyper parameters I choose for one model might not be ideal for the other model. For example, neural nets with relu activation function might produce the best results for classifying profit and loss, but logistic activation function might produce better results for classifying contribution per lot. Hence, as a machine learning practitioner a choice needs to be made regarding which activation function produces better results across all five models.

### 4.2.3 Accuracy Measures

Different accuracy measures were chosen for classification and regression techniques. For classification algorithms, it is important to note that for two categories: contribution per lot and clusters, multiclass classification was done out of three classes. For profit and loss, performance bonus and hard stops, binary classification was applied with two classes. The following accuracy metrics were taken for classification tasks:

- **Accuracy:** (true positives + true negatives) / total
- **Recall:** true positives / (true positives + false negatives)
- **Precision:** true positives / (true positives + false positives)
- **F1 Score:**  $2 \times ((\text{precision} \times \text{recall}) / (\text{precision} + \text{recall}))$
- **FBeta Score:**  $((1 + \text{beta}^2) * \text{Precision} * \text{Recall}) / (\text{beta}^2 * \text{Precision} + \text{Recall})$
- **Hamming Loss:** average loss or false labels per class in multi class classification

For Classification machine learning algorithms, samples in each bucket play an important role for predictions. For example, if the number of samples in a particular class are much lesser than the number of samples in other classes, then that particular class is much less likely to be predicted. This will produce



bias results against that class when the number of samples in that class in the testing or real-world's data are equally represented but are under-represented in the training set.

In Classification algorithms there is a trade-off involved between different accuracy measures (mentioned above) that we want to prioritize. Table 8 shows the differences and similarities between a model's predicted and actual results.

Table 8: Differences and similarities between predicted and actual results

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>

From the table 8, if Recall is prioritized (increase its value) and false negatives are reduced to make sure that we do not miss on any potential applicants that might end up being good traders, I face the risk of reducing Precision with it, as reducing false negatives can potentially lead to an increase in false positives as well. Hence, I have to balance the priorities accordingly.

For regression the following accuracy measures were chosen:

- **Mean Absolute Error:** average magnitude of the error, irrespective of direction.
- **Mean Squared Error:** average squared difference between the estimated values and the actual value.
- **R Squared:** evaluates the scatter of data points, higher the value, better it is.
- **Explained Variance Score:** explained variance regression score, best possible score is 1.0.

#### 4.2.4 Ethical Considerations for Model Selection

An important consideration in choosing the right machine learning model for an AI tool's predictions is also dependent on the transparency and explainability of the models. Similarly, transparency considerations can play an important role in evaluating the models if the organization developing an AI tool prioritizes ethics and transparency over other metrics like accuracy, resource requirements or time constraints of the development process. For the AI powered ed-tech tool developed for the financial services company, the final goal had been to empower the office managers and HR managers as 'humans in the loop' to take more informed decisions. Hence, it is essential for them to have some understanding regarding why the AI tool might be making certain predictions.

Algorithms like neural networks with hidden layers and tens of neurons in each layer are considered black box models with limitations on transparency. Similarly, support vector machines with 28 dimensions or features (in our case) are impossible for humans to perceive. Fig 18c shows svm Classification based on two features only. On the other hand, random forest algorithms are particularly useful for transparency and explainability in classification tasks. Using bootstrapped dataset with a sample of features to build trees, we can find the link between different features and model's predictions.

In real-world machine learning applications, classification is usually preferred over regression, as predicting a range of values gives machine learning algorithms a buffer for error, compared to predicting the exact value of a particular data point.

#### **4.2.5 Main Results**

The machine learning algorithms tested had differing performances across the five categories of data predicted. Tables 9a-d, 10a-d and 11a-d below show the results for all accuracy measures applied on classification and regression techniques for random forest, neural networks and support vector machines.

For contribution per lot classification, random forest with normalized data performed the best as shown by the Accuracy, Recall, Precision, F1 Score and Fbeta Score in table 9c. For contribution per lot regression svm seemed to perform better. For profit and loss classification, svm outperformed random forest and neural networks, irrespective of using normalized or non-normalized data. Though normalized data improved the accuracy of profit and loss classification predictions. For profit and loss regression svm seemed to perform better as well.

For Performance Bonus, random forest classification seemed to produce the best results, irrespective of using normalized or non-normalized data. For regression, svm produced better results, irrespective of using normalized data. For hard stop classification, neural nets with normalized data seemed to produce better results. For regression, svm produced better results, irrespective of using normalized or non-normalized data. For cluster classification, neural nets seemed to produce the best results. From the analysis, it seemed that in this particular company's context, normalized and non-normalized data do not have a huge impact on the accuracy of predictions.

The accuracy metrics values in classification and regression for each model and for every prediction category are given below.

For Classification in Contribution per lot, three classes were used with the following thresholds:

- Class 1  $\leq$  £0.1
- Class 2  $\leq$  £0.25
- Class3  $>$  £0.25

Tables 9a-d below show how different ML models (shown in section 4.2.2) performed with the Contribution per lot data in the normalized and non-normalized form.

Table 9a: Contribution Per Lot Classification with Non-Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	Fbeta Score	Hamming Loss
Random Forest Classification	0.404	0.397	0.379	0.382	0.379	0.600
SVM Classification	0.357	0.362	0.258	0.291	0.269	0.642
Neural Network Classification	0.404	0.397	0.414	0.399	0.406	0.595

Table 9b: Contribution Per Lot Regression with Non-Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	0.47	2.87	-0.16	-0.12
SVM Regression	0.40	2.53	-0.02	0.001
Neural Network Regression	0.48	2.55	-0.032	0.022

Table 9c: Contribution Per Lot Classification with Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	Fbeta Score	Hamming Loss
Random Forest Classification	0.45	0.45	0.44	0.45	0.44	0.55
SVM Classification	0.36	0.35	0.36	0.35	0.35	0.64
Neural Network Classification	0.36	0.35	0.35	0.34	0.35	0.64

Table 9d: Contribution Per Lot Regression with Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	0.47	2.9	-0.17	-0.13
SVM Regression	0.40	2.5	-0.03	-0.009
Neural Network Regression	0.41	2.5	-0.03	0.000

For contribution per lot classification, random forest with normalized data performed the best as shown in figure 9c. For regression SVM seems to perform better.

For Classification in Profit and Loss, we used two classes with the following thresholds:

- Class 1  $\leq$  £5000
- Class 2:  $>$  £5000

Tables 10a-d below show how different ML models (shown in section 4.2.2) performed with the Profit and Loss data in the normalized and non-normalized form.

Table 10a: Profit and Loss Classification with Non-Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
Random Forest Classification	0.45	0.45	0.45	0.45	0.45	0.55
SVM Classification	0.55	0.55	0.64	0.46	0.49	0.45
Neural Network Classification	0.5	0.5	0.25	0.33	0.28	0.5

Table10b: Profit and Loss Regression with Non-Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	7730.43	1.40	-0.26	-0.17
SVM Regression	6094.83	1.20	-0.082	-1.60e-06
Neural Network Regression	8883618.62	7.90e+13	-708918.92	-9.77e-15

Table 10c: Profit and Loss Classification with Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
Random Forest Classification	0.5	0.5	0.5	0.50	0.50	0.5
SVM Classification	0.60	0.60	0.60	0.60	0.60	0.40
Neural Network Classification	0.45	0.45	0.45	0.45	0.45	0.55

Table 10d: Profit and Loss Regression with Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	7743.41	1,40	-0.26	-0.17
SVM Regression	6094.78	1.20	-0.082	0.00
Neural Network Regression	9.25e+09	8.55e+19	-7.70e+11	1.22e-12

For profit and loss classification, svm seems to outperform random forest and neural networks, irrespective of using normalized or non-normalized data, though normalized data improves accuracy a bit. For regression, SVM seems to perform better as well.

For Classification in Performance Bonus, we used two classes with the following thresholds:

- Class 1  $\leq$  £500
- Class 2  $>$  £500

Tables 11a-d below show how different ML models (shown in section 4.2.2) performed with the Performance Bonus data in the normalized and non-normalized form.

Table 11a: Performance Bonus Classification with Non Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
Random Forest Classification	0.57	0.58	0.58	0.57	0.57	0.43
SVM Classification	0.52	0.53	0.54	0.51	0.52	0.48
Neural Network Classification	0.48	0.5	0.24	0.32	0.27	0.52

Table 11b: Performance Bonus Regression with Non Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	5333.71	74383038.41	-1.31	-1.02
SVM Regression	2899.07	37238798.81	-0.16	4.66e-06
Neural Network Regression	8.69e+23	7.55e+47	-2.35e+40	1.0

Table 11c: Performance Bonus Classification with Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
Random Forest Classification	0.57	0.58	0.58	0.57	0.57	0.43
SVM Classification	0.45	0.44	0.43	0.43	0.43	0.55
Neural Network Classification	0.38	0.37	0.36	0.36	0.36	0.62

Table 11d: Performance Bonus Regression with Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
<b>Random Forest Regression</b>	5331.62	74630707.17	-1.32	-1.02
<b>SVM Regression</b>	2899.04	37238548.19	-0.16	1.26e-05
<b>Neural Network Regression</b>	5694661.35	3.24e+13	-1009611.26	-9.77e-15

For Performance Bonus, random forest classification seems to produce the best results, irrespective of using normalized or non-normalized data. For regression, SVM seemed to produce better results, irrespective of using normalized data.

For Classification in Hard Stop counts two classes were used with the following thresholds:

- Class 1: = 0
- Class 2: > 0

Tables 12a-d below show how different ML models performed with the Hard Stop counts data in the normalized and non-normalized form.

Table 12a: Hard Stops Classification with Non Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
<b>Random Forest Classification</b>	0.52	0.52	0.52	0.52	0.52	0.48
<b>SVM Classification</b>	0.48	0.48	0.41	0.36	0.34	0.52
<b>Neural Network</b>	0.5	0.5	0.25	0.33	0.28	0.5

Table 12b: Hard Stops Regression with Non Normalized Data



Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	5.26	174.18	-0.45	-0.44
SVM Regression	2.55	124.3	-0.03	0.00
Neural Network Regression	5.27	120.60	-0.00	0.01

Table 12c: Hard Stops Classification with Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
Random Forest Classification	0.52	0.52	0.52	0.52	0.52	0.48
SVM Classification	0.60	0.60	0.60	0.58	0.59	0.40
Neural Network Classification	0.64	0.64	0.66	0.63	0.64	0.36

Table 12d: Hard Stops Regression with Normalized Data

Algorithm	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared (R2)	Explained Variance Score
Random Forest Regression	5.28	176.28	-0.46	-0.46
SVM Regression	2.46	125.43	-0.04	0.00
Neural Network Regression	6.28	133.03	-0.10	-0.07

For hard stop classification, neural nets with normalized data seemed to produce better results. For regression, SVM produced better results, irrespective of using normalized or non-normalized data.

For Classification in Clusters three classes were used with the following thresholds:

- Class 1 = Cluster 1
- Class 2 = Cluster 2
- Class 3 = Cluster 3 or Cluster 4

Tables 13a-d below show how different ML models performed with the Clusters data in the normalized and non-normalized form.

Table 13: Cluster Classification with Non Normalized Data

Algorithm	Accuracy	Recall	Precision	F1 Score	FBeta Score	Hamming Loss
<b>Random Forest Classification</b>	0.36	0.34	0.55	0.36	0.42	0.64
<b>SVM Classification</b>	0.45	0.45	0.47	0.46	0.46	0.55
<b>Neural Network Classification</b>	0.5	0.49	0.51	0.49	0.50	0.5

For cluster classification, neural nets seemed to produce the best results.

From the above findings it seems that in this company's context, normalized and non-normalized data do not have a huge impact on the accuracy of predictions.

#### 4.2.6 Conclusion

In the above context, I have five different models making predictions about traders' contribution per lot, profit and loss, performance bonus, hard stops and clusters. As expected, there is no single model that can produce the best results for all these five categories. In fact, changing a parameter in the model has different effects on different categories of predictions.

The choice of our model that is used in production is determined by not just the accuracy measures shown in the tables above, but also by the practical considerations. For example, random forest algorithms are usually preferred over neural networks for explainability and transparency. In the context of the AI-powered ed-tech tool for recruitment where we want to empower the 'human in the loop', explainability is crucial for enabling more informed decision making.

Furthermore, regardless of the specific technique used, increasing the number of observations the model is trained with is crucial to improve the accuracy of all models. Adding another set of data from a different modality, in addition to the survey constructs, to improve the prediction accuracy is being suggested in this report.

The tables above show that there were no significant differences between the accuracy metrics of these models for the five prediction categories. For some prediction categories, the random forest performed better, while for others neural networks or support vector machines out-performed random forest results.

For the second consideration: the explainability of the models, the random forest algorithms are known for their explainability and interpretability use cases (Petkovic et al, 2021; Fernandez et al, 2020; Neto and Paulovich, 2020; Vigil, 2016). Other algorithms like neural networks are considered black-box models (Zhang et al, 2018; Tzeng and Ma, 2005; Setiono, 2000; Rai, 2020). For the AI-powered ed-tech tool prepared for the financial services company, the explainability of the predictions was a necessary condition so the office managers who were using the tool could have a better understanding of why the tool was making certain predictions.

For the third consideration: the dataset prepared to train the machine learning models was not large enough with less than 150 datapoints, but the number of features was considerably better, with 28 features for each prediction category. In such instances, random forest models are reported to perform

better compared with neural networks or support vector machines (Wang et al, 2009; Muchlinks et al, 2016; Shaikhina, 2019). Hence, random forest algorithm was used for the AI-powered recruitment tool to make predictions.

The process described above, which led to the selection of the random forest algorithm for the AI-powered recruitment tool, has been documented in the Transparency Index Framework in the Machine Learning Modelling stage. This process also takes account of the decisions taken and assumptions made during the model selection stage of the AI tool development process. For example, this particular company wanted to reduce the false negatives as much as possible in the performance of the tool. They did not want the tool to underestimate the performance of a trader, even if there was a slight chance that he/she might excel at trading. This decision was incorporated in the machine learning modelling stage and documented in the model evaluation report and model card.

Considering the effectiveness of the model evaluation report in evaluating the choice of machine learning models, it was added as a requirement in the TIF for AI-powered ed-tech products. This report provides a framework to evaluate different ML models that AI practitioners may perceive to be suitable for a particular task. After preparing the model evaluation report and choosing a particular model, it is very important to go in-depth and evaluate the results and suitability of a particular model for the task in hand. The next section shows how a Model Card was used for this purpose in documenting the models used for the ed-tech tool.

### **4.3 Model Card**

Mitchell et al (2019) introduced a framework for reporting a model's details for AI-powered products in any domain. Their framework helps in documenting the details of the models used in AI products deployed in real world contexts. This model card was also prepared for the random forest algorithm used in the ed-tech tool. Next section shows the Model Card for the AI tool with the details of accuracy metrics for all prediction categories along with the ethical

considerations that were taken into account and some recommendations for improving the models' performance in future. There is some overlap between the data sheet shown above and the model card, but in essence, the data sheet focuses on documenting the quality of training data and model card focuses on documenting the performance of the machine learning model used in the AI development process. The different aspects of the data processing stage and machine learning modelling stage covered by these documents can play an important role in ensuring transparency for different stakeholders of an AI-powered ed-tech tool. Considering the popularity of these frameworks among the AI community to document their development processes (Garbin and Marques, 2022), both these documents were added as a requirement in the TIF along with some additional pointers to suit educational contexts.

The structure of the model card is as follows:

**4.3.1 Model Details** Basic information about the model

- This model was developed by the project lead.
- Model date: February 2020
- Model version: 1.0
- Model type and Info: Random Forest Classifier
- Information about training algorithms: Model was pretrained to predict traders' performance indicators and behavioural cluster shown in the table below:

Table 14: Recruitment tool's prediction categories and classes for candidates

Index	Prediction Category	Prediction	Classes		
1	Performance	Profit and Loss	Class 1 $\leq 5000$		Class 2 $> 5000$
2		Contribution Per Lot	Class 1 $\leq \pounds 0.1$	Class 2 $\leq \pounds 0.25$	Class 3 $> \pounds 0.25$
3		Performance Bonus	Class 1 $\leq \pounds 500$		Class 2 $> \pounds 500$
4		Hard Stop Counts	Class 1 = 0		Class 2: $> 0$
5	Behaviour	Clusters	Class 1 = Cluster 1	Class 2 = Cluster 2	Class 3 = Cluster 3 or Cluster 4

- Random forest classifier was chosen due to criteria such as the ability to scale to multi process implementation in future, its robustness to outliers, highly dimensional data and non-linear features, its ability to handle imbalanced classes, and also the bias-variance balance (i.e., each single decision tree is high on variance and low on bias, The averaging of all trees, keeps the bias low while moderating the variance). A detailed models' evaluation was conducted to analyse the results of different models like support vector machines and neural networks on the training data.
- Citation details: For modelling specifically we used python's open source library scikit learn from '[API design for machine learning software: experiences from the scikit-learn project](#)'
- License: This is an open-source software available at: <https://github.com/scikit-learn/scikit-learn>
- Where to send questions or comments about the model: Questions about this model can be directed to Educate Ventures Research.

**4.3.2 Intended Use:** Use cases that were envisioned during development.

- Primary intended uses: This model should only be used for predicting this particular company's traders performance and behavioural indicators mentioned in Table 14.
- Primary intended users: It should be used by office managers at this company for more informed decision-making when recruiting traders.
- Out-of-scope use cases: It should not be used for any domain other than trading and any company other than this.

**4.3.3 Factors:**

Factors could include demographic or phenotypic groups for which model performance may vary:

- The distribution of different factors that can potentially impact model's performance for certain groups of people have been shown in pie-charts in figure 13a-e above. These factors include gender, age, location and trading experience of traders whose data was used for training.

#### **4.3.4 Metrics:**

- Model performance measures: Model's performance measures include true positives, true negatives, false positives and false negatives. Models will be validated with traders' performance after around one year of joining the company. This is because training data consisted of traders who have been with this company for at least nine months.
- A set of co-design sessions were done by the project lead, discussing the various aspects behind each accuracy metric. As a result, from an ethical point of view, as well as the imbalance of class sizes, it was decided to tune the model such that the recall was prioritised, to try and minimise false negatives, specifically of the most 'weak' class, without hurting the recall of stronger classes, or alternative metrics.
- Decision thresholds: With 'human in the loop' approach used in this machine learning modelling implementation; decision thresholds are determined by office managers for each prediction category. Each prediction was accompanied by the probability of this prediction being correct when shown to office managers. The baseline decision thresholds are seen as being higher than the random model. That is 33% in case of three classes models, and 50% in cases of two classes models. This work was led by the project lead, and I was responsible for documenting its details in the Model Card.
- Variation approaches: With 'human in the loop' approach, tool was used in two different locations (for three different cohorts), tried on two different stages of the recruitment process, and shown to office managers in the form of a graphical user interface and an excel sheet. A models validation report was prepared by me based on how office managers perceived 'agreement' with the tool and 'high performance' of traders.

#### **4.3.5 Training Data**

Training data consisted of 140 instances with 28 features for profit and loss, performance bonus, contribution per lot and hard stop counts prediction

categories, and 72 instances with 28 features for clusters. All the information about training data is available in the Datasheet above in section 3.2.1.

#### **4.3.6 Quantitative Analysis**

- To overcome the problem of imbalanced classes, I considered methods such as oversampling and under sampling but recommended to use a two phased approach: first I and a project team member iterated with the domain experts on all kinds of options for binning the classes, such that the binning will make sense in terms of domain knowledge but will also be as balanced as possible. Then, I proposed handling the rest of the imbalances with class weighting. i.e., assigning higher weights in the models for minority classes, where needed. we thus compared balanced vs few schemes for weights for each model and chose what worked best.
- Project lead hyper tuned each of our five models, to adjust parameters such as depth, number of estimators, sample split and sample leaf. An exhaustive grid search was conducted, on all combinations of the parameters to achieve the best possible models.
- Normalizing the features was also experimented with all prediction categories. After experimenting with normalised and non-normalised features, I chose normalised features for clusters, contribution per lot and hard stop counts, and non-normalised features for profit and loss and performance bonus.
- For training the models, training data was divided into train and test datasets with around 7:3 ratio, we also tested using cross validation, although for random forest it was potentially less effective. Details of the accuracy measures of different classes for each prediction category is given in the tables in section 4.2.5.

#### **4.3.7 Ethical Considerations (as specified in the Model Card)**

A detailed auditing plan was used for developing the AI tool. Several measures were taken to make the tool explainable and more understandable



for the office managers who would be using it. Python’s package Lime was used to add explainability to the models so ‘humans in the loop’ or office managers can see why the model is making a particular prediction. Training sessions were also conducted with office managers to help them in evaluating models’ predictions before they used it with candidates.

#### 4.3.8 Caveats and Recommendations

The project team’s recommendations for improving the models are as follows:

- Retraining the models with more data from 2019 and 2020 traders;
- Adding multi-modality to enrich the training data;
- Training office managers after taking account of the findings from the model validation report (prepared by me) and recruitment tool’s alpha version evaluation interviews that were conducted by the project lead.

Tables 15a-e below illustrate the evaluation measures of different classes in each prediction category. For profit and loss, the accuracy of the model (used in production) was 0.55 which means the model accurately predicts 55% of the time. This metric is calculated by adding the true positives to true negatives and dividing the total number of predictions made (true positives plus true negatives plus false positives plus false negatives).

Table 15a: Accuracy Measures of two classes in Profit and Loss predictions

No of Classes	Precision	Recall	F1-Score	No of Predictions
Class 1 <= 5000	0.5	0.68	0.58	19
Class 2 > 5000	0.62	0.43	0.51	23

For performance bonus the accuracy of the model (used in production) was 0.40 and the performance of the model across various evaluation metrics for different classes is shown in the table 15b below.

Table 15b: Accuracy Measures of two classes in Performance Bonus predictions

No of Classes	Precision	Recall	F1-Score	No of Predictions
Class 1 <= 500	0.39	0.67	0.49	18

Class 2 > 500	0.45	0.21	0.29	24
---------------	------	------	------	----

For hard stop the accuracy of the model (used in production) was 0.53 and the performance of the model across various evaluation metrics for different classes is shown in table 15c below.

Table 15c: Accuracy Measures of two classes in Hard Stop predictions

No of Classes	Precision	Recall	F1-Score	No of Predictions
Class 1 = 0	0.55	0.72	0.62	25
Class 2 > 0	0.50	0.32	0.39	22

For contribution per lot the accuracy of the model (used in production) was 0.38 and the performance of the model across various evaluation metrics for different classes is shown in the table 15d below.

Table 15d: Accuracy Measures of three classes in Contribution Per Lot predictions

No of Classes	Precision	Recall	F1-Score	No of Predictions
Class 1 <= 0.1	0.36	0.72	0.48	18
Class 2 <= 0.25	0.57	0.29	0.38	14
Class 3 > 0.25	0.25	0.07	0.11	15

For clusters the accuracy of the model (used in production) was 0.54 and the performance of the model across various evaluation metrics for different classes is shown in the table 15e below.

Table 15e: Accuracy Measures of three classes in Clusters predictions

No of Classes	Precision	Recall	F1-Score	No of Predictions
Class 1 <= 0.1	0.60	0.86	0.71	7
Class 2 <= 0.25	0.20	0.33	0.25	3
Class 3 > 0.25	0.67	0.43	0.52	14

## 4.4 Conclusion

Model Cards provide an effective framework to document the different components of a machine learning model. Hence, they are suggested as a requirement in the Transparency Index Framework for ‘increasing transparency into how well (machine learning modelling part of) artificial intelligence technology works’ (Mitchell et al, 2018). To increase the suitability

of Model Cards for educational contexts, some additional requirements for the reasoning behind the model choice were also added to facilitate the needs of different stakeholders of AI in education. For example, to what extent transparency and explainability considerations are taken into account when choosing a particular model, or which tools are used to make the ML models more understandable for different stakeholders?

ML modelling is often considered the backbone of an AI system. It is responsible for producing the decisions or predictions made by the AI system. Models Evaluation Report and Model Cards can be considered an integral part of transparency considerations to make this ML modelling stage of the AI development process transparent. Models Evaluation Report (as shown in section 4.2) document and systematise the machine learning model selection process for an AI system and Model Cards (as shown in section 4.3) document the details of the machine learning model's performance including the model evaluation and testing procedures used, hyperparameter optimisation techniques utilised and other decisions or assumptions made in the process. Hence, they are suggested as requirements for transparency in the TIF.

After ML modelling predictions, AI systems are deployed in the real world. The work on AI systems does not stop after deployment. The next chapter shows the different transparency considerations that were taken into account during the deployment and iterative improvements stage of the AI-powered ed-tech tool developed for the financial services company.

# Chapter 5: Phase 2 - Framework Creation: Deployment and Iterative Improvements Stage

## 5.1 Introduction

When an AI product or tool has passed through the data processing and machine learning modelling stages in its development, it should be ready to be deployed in the real-world for users. At this stage, it is very important to monitor and evaluate the performance of the tool to ensure that it works as expected. Deployment is not straightforward, there can be significant differences between the training data on which an AI system has been trained, and the real-world data which the AI system must process after deployment as discussed in section 2.8.4. Researchers at IBM have developed technical tools such as AI Fairness 360 (Bellamy et al, 2019) that evaluate models as well as datasets according to a certain fairness metric. This includes unwanted bias in the training data that 'places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage' over others (Bellamy et al, 2019). Hao and Stray (2019) have shown how the COMPAS algorithm used in the US judicial system was biased against one particular group compared to another.

The potential negative impacts of AI-powered ed-tech products mean that the deployment and iterative improvement stages play a crucial role in the success of AI systems in educational environments. A single wrong prediction, for example, that a student may drop out from the course of study in the near future, can change the teacher's and parents' attitude towards that student and could lead to devastating psychological impact. Similarly, a correct prediction from an AI system that a student is at the risk of dropout in near future could enable educators and parents to intervene and help the learner.

The predictions of AI-powered ed-tech tools in HR contexts can directly impact the manager's decision to recruit or not to recruit a candidate. A wrong prediction can mislead the OMs and lead to major financial losses or gains for

the company through the quality of candidates they recruit. For AI-powered ed-tech products, it is very important to first test them in different contexts before deploying them in the real-world at scale. In the data processing and machine learning modelling stages of the AI development process, the strengths and weaknesses of the data and ML models used should have been identified. These pros and cons can be either further increased or mitigated, depending on how the tool is deployed in the real-world. For example, a learning analytics dashboard that visualizes the learning progress of different students in a classroom may have been designed and tested in classrooms with one teacher and less than twenty students in an independent fee-paying school in a particular country and a specific area, where the vast majority of students are white. Such a dashboard may not therefore work as effectively when deployed in a classroom in a state funded school with two teachers (an additional teaching assistant) and more than thirty students, who are a mix of different ethnic backgrounds. Hence, transparency is important to ensure the expected performance and fairness of AI systems in educational contexts. The ed-tech tool developed for the financial services company faced similar issues as the data collected from some locations was a lot more compared to other locations.

To mitigate these imbalances, when an AI tool is deployed in the real-world, it should be validated to ensure that it is performing as expected. These considerations were taken into account when the AI-powered ed-tech tool was built for the financial services company as discussed in section 5.2 below. A model validation report was prepared after conducting interviews and taking thorough feedback (in an Excel sheet) from the office managers who used the alpha version of the AI tool. Four office managers were interviewed, two from Poland and Ukraine and two from India. For all the Office Managers (OMs), English was their second language, but they were comfortable communicating in English. These interviews were conducted by the project lead but some of the findings from the interviews were added to the model validation report that I prepared.

OMs were also asked to share a spreadsheet highlighting the decisions they took for different candidates and if they thought their decisions were in agreement with the AI tool or not. I prepared the model validation report by comparing this data provided by the office managers with the actual predictions provided by the ed-tech tool on profit and loss, contribution per lot, performance bonus and hard stop counts. The model validation report illustrated in table 17 shows the details of how the tool performed with three different cohorts that were recruited in three different locations: Ukraine, Poland and India. It also showed how office managers from different locations who were acting as the 'human in the loop' and a learner from the AI tool, perceived the predictions produced by the AI. The model validation report in the next section also confirmed that the perceived learning of office managers from the AI-powered ed-tech tool was different in different contexts. For example, for the Polish and Ukrainian office managers the design of the ed-tech tool was more effective in presenting the AI-powered predictions, and in evaluating the candidates, compared to Indian office managers. The agreement of these office managers with tool's predictions was much higher compared to office managers in the Indian office.

## **5.2 Model Validation Report**

### **5.2.1 Introduction**

The purpose of the validation report was to evaluate the AI tool's models, strictly in terms of their predictions and their alignment with Office Managers (OMs) judgement and expectations. At the time when this report was prepared, I was not able to directly compare the predictions of the models used in the AI tool with real world data, because the training data consisted of traders that had been working for the company for at least nine months. Therefore, the next phase of validation required a few months (around nine months) to gather enough live trading data from new traders for whom the AI tool's predictions were used.

The AI tool's models are not evaluated based on their performance in general, but in the context of the decisions of the 'human in the loop' i.e., OMs judgements. Currently, these models are evaluated based on their alignment

with the decisions of the OMs. This data was shared by the OMs with me in the form of an excel sheet.

After at least nine months of live trading by the traders who were evaluated using the AI tool, the validation report would be updated by comparing the tool's predictions and the office manager's judgement with actual trading data from those traders who were hired. This will still give a partial image of validating the models, since I do not have data on the potential performance of those who were not recruited. The aim, however, is to constantly evaluate and validate the models based on the evidence we can apply.

### **5.2.2 Configurations of the tool**

The AI tool was used in three different configurations in the period from March to April 2020, just before the COVID pandemic broke, which halted the anticipated further use of the tool. The three configurations are different in terms of three main conditions:

Configuration 1. The format in which the predictions were presented to the OMs was not in the designed point of decision making (after stage 4 of the recruitment process before the assessment day). The designed context included data about the scoring across all of the recruitment stages which included application forms, questionnaires, math tests, videos submitted by candidates and assessment day. The out-of-context setting was an Excel report including just the probability of each prediction, with limited explanations.

Configuration 2. The predictions were presented to the OMs at their designed point of decision making. The designed point of decision making was after stage four, to aid with the decision of sending candidates to the assessment centre, taking into account that the assessment centre is a very expensive and human-intensive stage.

Configuration 3. The number of candidates to be hired in a particular office's recruitment drive. There were three different offices responsible for recruiting three different cohorts of traders. The number of candidates hired were

compared with the number of applicants that were selected in stage four or stage five, depending on for which recruitment stage the tool was used. The percentage of candidates hired or invited to the assessment centre could also influence the evaluation of the tool.

Table 16 highlights the different configurations in which these predictions were used by OMs in different offices. For the trading roles, OMs reviewed 132 applicants from the Indian cohort, 63 applicants from the Polish cohort and 25 applicants from the Ukrainian cohort. The percentages in tables 16 and 18, and figure 19 were calculated from the available data (that is 115 Indian applicants, 59 Polish applicants and 22 Ukrainian applicants) that was shared by the office managers in the form of an Excel sheet.

In addition to the difference in the independent setting conditions, the predictions were all experimental, the different OMs used the predictions in slightly different ways (according to different configurations above). It was anticipated that these differences would have had a huge influence on the OMs eventual use and added value by the AI tool. These differences are also documented in Table 16. It is important to note that the Polish OMs used the tool as a confirmatory tool throughout (the recruiting of Polish and Ukrainian candidates), whether it was used before or after the assessment centre stage.

Table 16: Summary of how the AI tool was used in different offices for decision making

<b>Cohort of traders</b>	<b>Predictions presented in designed context (with the five stages' data)</b>	<b>Predictions were presented in their designed point of decision making</b>	<b>Tool's Contribution</b>	<b>Prediction categories taken into account</b>	<b>Percentage of candidates recruited</b>
<b>Polish (April 2020)</b>	Yes	Yes	Tool was used to validate the initial decision of inviting the candidate for assessment centre or not	Profit and Loss, Performance Bonus and Clusters	77%
<b>Ukrainian (April 2020)</b>	Yes	No, the tool was used after candidates had gone through the assessment centre (stage 5) of	Tool was used to validate the conclusions from the whole recruitment	Profit and Loss, Performance Bonus and Clusters	46%



		the recruitment process	process (stage 1 to 5)		
<b>Indian (March 2020)</b>	No	No, the predictions alone were used after candidates had gone through the assessment centre (stage 5) of the recruitment process	Tool was used to decide between competing top candidates, as a tie breaker after stage 5	Profit and Loss, Performance Bonus, Contribution Per Lot, Hard Stop Counts and Clusters	9%

The AI tool predicted the class of four performance features (profit and loss, contribution per lot, hard stop counts and performance bonus) and one behavioural feature (clusters) that an applicant would belong to, as shown in Table 17.

The focus of interest is not on the predictions of behavioural clusters of candidates as they are not directly related to their performance as traders. The focus is on the classification of performance metrics of traders and their interpretation by OMs, although it was known that the predictions of clusters were taken into account in some of the offices, at least to some extent.

### 5.2.3 Analysis

During the study, it became clear to me that the semantics attached to what is described as a ‘high performance trader’ is not identical between the three different offices. In addition, what the OMs referred to as the criteria which defined ‘agreement with the tool’ were not the same. This use of the ed-tech tool highlights the need to discuss standardization across hiring practices among the various offices of the company.

Table 17 summarises the models considered in the hiring decisions for each office, their subjective definition of what is ‘high performance, and which criteria they seem to have adopted to serve as sufficient for an ‘agreement with the tool’. Tables 19 (a-d), 21 (a-d) and 22 (a-d) below provide details of the distribution of different classes within each prediction category for the candidates that were accepted or rejected by OMs in different offices.

Table 17: Tool's Predictions and Office Managers evaluation of candidates

Cohort	Prediction categories taken into account as claimed by Oms	What is 'high performance'	What is agreement with the tool
<b>Polish (April 2020)</b>	<ul style="list-style-type: none"> <li>Profit and Loss,</li> <li>Performance Bonus • Clusters</li> </ul>	Profit and Loss = Class 2, Performance Bonus = Class 2, Contribution Per Lot = Class 1, Hard Stop Counts = Class 2	Profit and Loss and Performance Bonus predictions are in a higher class i.e., class 2. At times one of them in a higher class would also work
<b>Ukrainian (April 2020)</b>	<ul style="list-style-type: none"> <li>Profit and Loss,</li> <li>Performance Bonus • Clusters</li> </ul>	Profit and Loss = Class 2, Performance Bonus = Class 2, Contribution Per Lot = Class 1, Hard Stop Counts = Class 1	Profit and Loss and Performance Bonus predictions are in a higher class i.e class 2. At times one of them in a higher class would also work
<b>Indian (March 2020)</b>	<ul style="list-style-type: none"> <li>Profit and Loss,</li> <li>Performance Bonus</li> <li>Contribution Per Lot</li> <li>Hard Stop Counts</li> <li>Clusters</li> </ul>	Profit and Loss = Class 2, Performance Bonus = Class 2, Contribution Per Lot = Class 1, Hard Stop Counts = Class 2	Profit and Loss and Performance Bonus predictions are in a higher class. At times one of them in a higher class would also work

Table 17 shows that the analysis of predictions is subjective to each OM's interpretations of the prediction categories and their contexts. Some OMs may use all prediction categories unlike others as shown in column 1.

Unfortunately, there was not enough data from the platform logs about what prediction categories they used. The table shows that different OMs focused on different sets of prediction categories. The second column of Table 17 shows different classes for each prediction category for candidates that were categorised as 'high performance' by OMs or were accepted. In the third column of Table 17, the analyses of data about candidates is presented who were accepted and explore how OMs were defining 'high performance' and 'agreement with the tool'.

According to the excel sheet that OMs shared, at times they accepted the candidates as 'high performance' even if one of profit and loss or performance bonus, had a lower predicted class. Table 18 illustrates different types of agreements for the candidates that were hired/invited by OMs. In this table OMs' agreement with the tool is highlighted in three categories that were derived from the data analysis of the Excel sheet that OMs shared:

- Agreement with the tool, as perceived by OMs. These numbers about the agreement with the tool were provided by OMs. For example, Polish OMs stated that for all the candidates they considered for the Polish cohort, 86% of the time they believed their judgement was in agreement with the AI tool because its predictions were high for those candidates (as seen in table 18).
- Actual agreement (or alignment) with the tool defined as a higher class for two prediction models, Profit and Loss **and** Performance Bonus (i.e., the percentage of candidates predicted high in PnL, high in bonus and were also selected by OMs, plus candidates not predicted high in PnL and high in bonus who were not selected by OMs);
- Actual agreement (or alignment) with the tool defined as a higher class for the Profit and Loss prediction category only.

Table 18: Office managers' different types of 'agreements' with the tool

<b>Cohort</b>	<b>Agreement as stated by Office Managers</b>	<b>Agreement taken as a higher class for both Profit AND Loss and Performance Bonus</b>	<b>Agreement taken as a higher class for either Profit and Loss OR Performance Bonus</b>	<b>Agreement taken as a higher class for Profit and Loss only</b>
<b>Polish (April 2020)</b>	86%	59%	79%	65%
<b>Ukrainian (April 2020)</b>	54%	17%	67%	42%
<b>Indian (March 2020)</b>	28%	16%	38%	22%

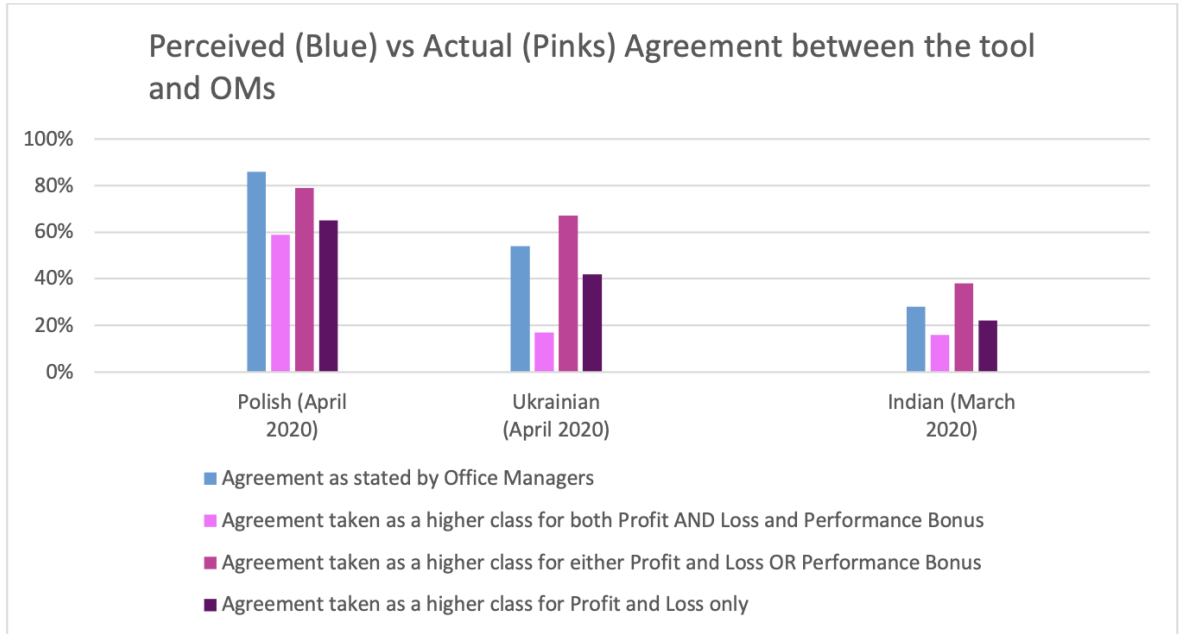


Figure 19: Perceived (Blue) vs Actual (Pinks) Agreement between the tool and OMs

Table 18 and Figure 19 illustrate that OMs’ perceived agreement with the AI tool can vary across different offices and is significantly different to the actual agreement with the tool. However, OMs’ perceptions about their agreement with the tool are not completely random, rather a number of factors may influence how the OMs as ‘humans in the loop’ evaluate agreement with the algorithm’s predictions. For example, data for the Polish cohort included in Table 18, indicate that agreement with the tool for the candidates invited to the assessment centre varies according to different interpretations of ‘agreement’. OMs’ personal evaluation or intuition about the candidates affects (or even biases) their interpretation of ‘agreement’ with the tool. For instance, two candidates with very similar predictions, such as higher class for profit and loss and lower class for all other categories, might be evaluated differently with respect to agreement with the tool because an OM’s intuition about one candidate was very positive compared to the other. This can be explored further in the future work.

It is also important to highlight the complexity of evaluating the contexts of different applications of the recruitment tool. If the tool is used after stage 5 in the recruitment process, then the ground truth will be whether the OMs hired

the candidate or not. But, the tool does not predict whether the candidate should be hired. It only predicts certain performance and behavioural indicators as illustrated in [Table 14](#). Next, the context for each office is discussed under their respective headings.

### 5.2.3.1 Indian Office

The data available for the Indian office was predictions data of 10 out of 12 candidates who were recommended for hire. The data for candidates who were not hired was for around 105 traders (variable across different prediction categories) out of 120 traders. The distribution of different prediction categories is given in the Tables 19 a-d:

Table 19a: Performance Bonus classification for Indian cohort

	Class 1	Class 2
Hired	30%	70%
Not Hired	17%	83%

Table 19b: Profit and Loss Classification for Indian cohort

	Class 1	Class 2
Hired	30%	70%
Not Hired	17%	83%

Table 19c: Hard Stops classification for Indian cohort

	Class 1	Class 2
Hired	30%	70%
Not Hired	54%	46%

Table 19d: Contribution Per Lot classification for Indian cohort

	Class 1	Class 2	Class 3
Hired	100%	-	0%
Not Hired	91%	-	9%

Indian OMs did not use the tool with a graphical user interface. They received predictions in excel sheets and compared the models' predictions with their own judgement after candidates had gone through the assessment centre. They used these predictions after stage 5 of the recruitment process. The figures in Table 20 regarding the candidates hired or not hired and predictions matching with OMs' judgment or not, were provided by OMs. For example, 11 candidates who were predicted as 'high performance' by the tool were also hired by the office managers.

Table 20: Alignment of Recruitment Tool's Models with Indian OMs as perceived by them

Office managers hired/invited the candidates predicted as 'high performance'	Office managers did not hire/invite the candidates not predicted as 'high performance'	Office managers did not hire/invite the candidates predicted as 'high performance'	Office managers hired/invited the candidates not predicted as 'high performance'	Inconclusive
11	22	30	1	68

Indian OMs neither reported agreement nor disagreement with the tool in the case of the inconclusive candidates included in Table 20.

### 5.2.3.2 Polish Office

The data for the Polish office was predictions data for 44 out of 45 candidates who were invited to the assessment centre and 15 out of 18 candidates who were not invited.

The distribution of classes between different prediction categories is given in the Tables 21a-d: For every prediction class, these tables show how many candidates were invited or not invited to the assessment day in Poland. For example, in table 21a, among all the candidates who were invited by OMs in Poland, 28% were predicted in class 1 by the tool.

Table 21a: Performance Bonus classification for Polish cohort

	Class 1	Class 2
Invited	28%	72%
Not Invited	43%	57%

Table 21b: Profit and Loss Classification for Polish cohort

	Class 1	Class 2
Invited	16%	84%
Not Invited	43%	57%

Table 21c: Hard Stops classification for Polish cohort

	Class 1	Class 2
Invited	32%	68%
Not Invited	29%	71%

Table 21d: Contribution Per Lot classification for Polish cohort

	Class 1	Class 2	Class 3
Invited	84%	-	16%

Not Invited	86%	-	14%
-------------	-----	---	-----

The predictions data of all candidates in the Ukrainian cohort, who were hired was available and that for 13 out of 16 candidates who were not hired.

The distribution of classes between different prediction categories is given in the Tables 22a-d. For every prediction class, these tables show how many candidates were hired or not hired in Ukraine. For example, in table 22a, among all the candidates who were hired by OMs in Ukraine, 22% were predicted in class 1 of the Performance Bonus metric by the ed-tech tool.

Table 22a: Performance Bonus classification for Ukrainian cohort

	Class 1	Class 2
Hired	22%	78%
Not Hired	30%	70%

Table 22b: Profit and Loss Classification for Ukrainian cohort

	Class 1	Class 2
Hired	22%	78%
Not Hired	20%	80%

Table 22c: Hard Stops classification for Ukrainian cohort

	Class 1	Class 2
Hired	83%	17%
Not Hired	88%	12%

Table 22d: Contribution Per Lot classification for Ukrainian cohort

	Class 1	Class 2	Class 3
Hired	78%	-	22%
Not Hired	40%	-	60%

Polish OMs used the tool's graphical interface, with explanations for Ukrainian and Polish cohorts. For Polish candidates the tool was correctly used to decide if a particular candidate should be invited to the assessment centre or not, after stage 4 – video submission. For Ukrainian candidates, the AI tool was used after stage 5, mostly to decide between the competing top candidates.

The figures in Table 23 and 24 regarding the candidates hired or not hired and predictions matching with OMs' judgment or not, were provided by OMs. These figures reflect on the AI-powered ed-tech tool's performance in

facilitating office managers to invite or not to invite the candidates to the assessment day.

Table 23: Alignment of Recruitment Tool's Models in Polish Office for Polish Candidates, as perceived by OMs

Office managers hired/invited the candidates predicted as 'high performance'	Office managers did not hire/invite the candidates not predicted as 'high performance'	Office managers did not hire/invite the candidates predicted as 'high performance'	Office managers hired/invited the candidates not predicted as 'high performance'	Inconclusive
43	11	1	2	6

Table 24 (below): Alignment of Recruitment Tool's Models in Polish Office for Ukrainian Candidates, as perceived by OMs

Office managers hired/invited the candidates predicted as 'high performance'	Office managers did not hire/invite the candidates not predicted as 'high performance'	Office managers did not hire/invite the candidates predicted as 'high performance'	Office managers hired/invited the candidates not predicted as 'high performance'	Inconclusive
9	4	5	0	7

OMs reported that the confidence of the tool was less than 55% for two class predictions for inconclusive candidates, as illustrated in the last columns of Tables 23 and 24.

Tables 23 and 24, and Figure 19 illustrate that the AI tool's models seem to better align to decisions made for Polish candidates than Ukrainian candidates. This might be attributed to the context of the tool's usage, or to the percentage of candidates hired or the demographics of training data which mostly consisted of traders based in Poland (43%). The alignment of the tool's suggestions and recruitment decisions on Ukrainian candidates is nevertheless, better than for Indian candidates. These differences could relate to the cultural similarities between Ukraine and Poland, compared with the huge difference (context-less) in the way OMs used the tool for Ukrainian and Indian candidates, as showed in Table 16.

The alignment of the AI tool's suggestions and an OM's final recruitment decision is much lower with Indian candidates, compared to Polish and Ukrainian candidates. This can potentially also be attributed to the high number of candidates not hired who were not strongly predicted by the tool



and the high number of candidates not hired, who were strongly predicted by the tool, or to the under representation of Indian candidates in the training data (2% only). The real reasons behind these differences in different offices can only be confirmed after the tool has been applied in a uniform context across offices, with minimal differences in the percentages of candidates hired.

#### **5.2.4 Discussion**

This chapter has presented a preliminary analysis of the validation of machine learning models used for educating OMs through the predictions of traders' trading performance at a financial services company. Models were validated by indirectly comparing predictions about traders' performance with the OMs' perceived and actual agreements with the tool. The analysis presented shows that typically the definition of a high performing trader varies across offices. In addition, the OMs definition of what 'agreement' with the tool means, is different across different managers. OMs perceived agreement is closer to the 'profit and loss' or 'performance bonus' condition. There were instances where traders' profit and loss, as well as performance bonus was predicted to be high, but they were rejected by OMs and still considered as an agreement with the tool. A possible explanation might be confirmation bias on the part of OMs or that OMs thought a candidate could be a good trader but lacked the language skills or was aggressive/toxic. They may reject the candidate for reasons other than performance.

The gaps between the perceived and the actual agreement with the tool point to another, unaccounted for component in the decision making of the OMs. This implicit component may be based on their subjective impression about the candidate, their own experience, their cognitive biases (e.g., anchoring bias, availability heuristic or confirmation bias), or a range of other factors that might make them interpret the prediction as confirmatory to their pre-made decision.

The AI-powered ed-tech tool discussed in this chapter was developed and designed with the aim of preserving the agency of the human decision

makers, by empowering and educating them with useful information, rather than making the decision for them. It was not a recruitment tool responsible for screening or selecting potential candidates, but an ed-tech tool enabling OMs to take more informed decisions. It is interesting to note that from conversations with the OMs (that were conducted by the project lead), a conflicting narrative was clearly shown. On the one hand, they do not want to let the tool make the decision for them, they question its validity, and are slow to trust it. On the other hand, they are keen to get more decisive support from it. The fact that the tool stops a few steps away from the actual and final decisions, and gives five different predictions, put the OMs in a position where they could not get 'a clear answer' from the tool, which sometimes precipitated a frustrated reaction. The tool does not tell them whether or not to hire a trader, it does not tell them whether the trader will become a 'high performing trader', and not even what a 'high performing trader' is.

However, OMs were expected to integrate the predictions, and the other contextual data provided by the tool, with their own impressions and intuitions, before making their final decision. The tool did not make the hiring decisions for OMs, but only educated them with an extra piece of information regarding the applicants' potential future performance as traders. It is clear from the analysis presented here, that the degrees of freedom offered by the tool, have major implications for its interpretation and use as a decision support ed-tech tool.

The OMs are faced with a range of information about candidates during the recruitment process. Some of this information is produced by the tool, and some is collected through each OMs' own sensory, emotional and cognitive systems. When each OM reaches the final step of making a decision, they might take into account any myriad of such evidence, as researchers there is no access to all the information available to OMs. The gap between the perceived and actual agreement of OMs as illustrated in this chapter suggests that the OMs have an inability to clearly separate between the different types of information available to them: the evidence they were given by the tool, and

the evidence they derive from a range of other sources that becomes implicitly joined with the evidence given by the tool.

Another important conclusion relates to the difference in agreement levels (both perceived and actual) between the offices. As shown in Table 18 and Figure 19, the Indian OMs, who used the tool without the contextual data and in a recruitment stage different from that for which the tool was designed, showed the lowest degree of agreement. The polish OMs, who used the tool twice, clearly showed higher levels of agreement (both perceived and actual) when using the tool at the designed recruitment stage.

What is clear from the analysis reported here is that there is evidence that *using the AI tool as designed*, meaning both the designed contextual data and the designed recruitment phase, produced higher agreement levels between the tool and the OMs.

### **5.2.5 Limitations and future investigations**

Chapter 5 presented a comparison between machine learning models predicting the classes for different performance and behavioural indicators of traders, and humans (OMs) viewpoint on those predictions. The tool did not directly predict if the candidate should be hired or not. This gap between the tool's predictions and OMs decision to hire or not to hire is filled by the OMs own judgment. Their judgment is also influenced by the context for the tool's application as discussed above. This makes it an ed-tech tool rather than a recruitment one. The tool was not recruiting candidates or providing recommendations, but the purpose of the AI tool was to provide an extra source of information to OMs when they are recruiting candidates. This information was based on the predictions of the metrics they used for evaluating traders' performance.

In terms of traders' performance and behaviours, a direct validation of the models should be conducted when there is sufficient trading data from applicants who were hired. Unfortunately, there is no feasible way to evaluate performance and behaviour of traders who were not hired, because we cannot access their trading data, if indeed they did pursue a career in trading.

To collect more data for the models' validation, it is recommended to add a filter that enables OMs to choose the prediction models they want to see. This would enable the collection of data about the prediction categories that have the biggest impact on the OMs interpretation of the tool's predictions. Also, it would be useful to collect some clickstream data about whether and how OMs used the predictions' explanations available through the tool's interface.

The model validations report should be updated in a few months by comparing the models' predictions with real world data - applicants' actual performance as traders after they were hired. However, it should be noted that there will still be a gap between the tool's predictions and the OMs' decision. This gap is filled by each OM's interpretation of the predictions and is influenced by the context in which the tool is used.

The findings from the model validation report of the tool showed that there were differences between how OMs from different locations who used the tool. They also show that OMs have different perceptions of high performing traders and there were gaps between what OMs perceived as agreement with the tool and what was actual agreement with the tool. On one hand, OMs did not want the tool to have complete decision-making authority and questioned its validity, but on the other hand they wanted more decisive support from the AI tool. These findings from the model validation report would play an instrumental role in improving the future iterations of the AI-powered ed-tech tool.

### **5.3 Conclusion**

The model validation report discussed in this chapter was added as a requirement in the Transparency Index Framework, because it offers a very useful framework to test AI-powered ed-tech products in different contexts. After an AI tool is deployed in an educational setting, it is very important to observe and evaluate the tool's impact. Is it performing as expected? How does it impact on the environment of the educational setting? How does the context of tool's deployment impact on its performance and perception among users?

How does it affect the teachers, learners, the communication between teachers and learners and the communication among learners? Such transparency investigations on the impact of the ed-tech would enable ed-tech companies developing machine learning powered AI tools to keep evidence at the heart of their product development.

### **5.3.1 First version of the Transparency Index Framework**

The first version of the framework that was prepared after the literature review and application of several ethical AI frameworks in developing an AI-powered ed-tech tool is given below:

Transparency for different stages of an AI tool development process:

- Data Processing Stage:
  - How was data gathered?
    - What were its sources?
    - Was informed consent given from all individuals?
  - How was data normalized?
  - What techniques were used to process the data?
  - What data on sensitive variables is collected?
  - How was the sensitive variables data processed and stored?
  - What types of biases were identified in the data:
    - Historical Bias,
    - Representation Bias,
    - Measurement Bias,
    - Aggregation Bias,
    - Evaluation Bias,
    - Deployment Bias.
  - Are these biases being shown to humans in the loop when they see the AI system's predictions?
  - Did the 'humans in the loop' receive any training on how to interpret AI system's predictions?
- Machine Learning Modelling Stage:
  - Which ML model was used for predictions?

- Why was this particular model chosen? Was any experimentation done with different models?
  - In choosing the model were Transparency capabilities of the model taken into consideration?
- Is the model doing regression or classification?
- Is the model using any Explainable Artificial Intelligence (XAI) tools or providing explanations of the predictions:
  - If yes, which XAI tools are being used?
  - What are the strengths and weaknesses of these tools?
    - Was 'human in the loop' trained regarding the limitations of these explanations?
    - Were any measures taken to address these limitations in autonomous AI systems?
- How was the machine learning model audited? For example, what were the results of using counterfactuals etc.
- Deployment and Iterative Improvements Stage
  - What security and privacy measures were taken when deploying the AI system?
  - How many people will be directly impacted by this AI system?
  - How many people will be directly involved in deploying this system?
  - Does this AI system come with a Disclaimer section in the form of text highlighting the contexts in which this AI system cannot be used?
  - Is there some form of visual signaling to indicate that aspects of this AI system are work in progress, or are not perfect or have certain biases against these particular groups?
    - What is the carbon footprint of training the ML models being used in this AI system? For example, some cloud providers provide information on environmental impact of training a particular ML model.

There are a number of factors that influence the type of transparency that should be induced in an AI system. Some of these factors are as follows:

- The system is autonomous or has a 'human in the loop'.
- The tech savviness of individuals using the tool.
- The tech savviness of individuals impacted by the tool.
- Legal obligations in the geographic location where the tool is deployed.
- Type of Machine Learning technique or model being used.
- Tech infrastructure on which ML is implemented, for example batch processing or stream processing in real-time.

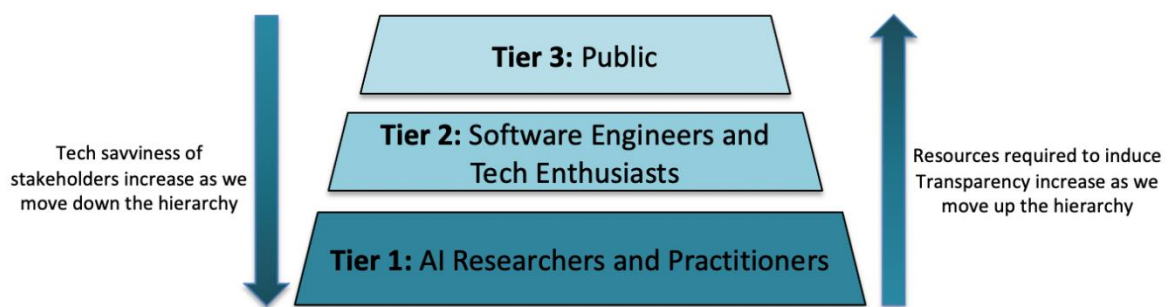


Figure 20 (above): Three Tiers of Transparency

Figure 20 shows how the different users of an AI system in educational contexts can be classified into three groups based on their tech savviness. These three categories of stakeholders were identified during the knowledge elicitation process with the domain experts, office managers who were responsible for the recruitment (and were the direct users of the AI tool) and technical staff such as software engineers who helped in the tool's deployment.

It was noted that different measures are required for making the AI-powered ed-tech tool transparent for the stakeholders. These measures are dependent on the tech savviness of the stakeholders and the resources required to make the tool transparent for different stakeholders also vary. Table 25 in the framework below shows the measures needed to make an ed-tech tool transparent for the three tiers.

Table 25 (below): Three Tiers of Transparency

Tiers	Description
Tier 1	<p>An AI system is considered Transparent for AI researchers and practitioners if they know or have access to the:</p> <ul style="list-style-type: none"> <li>• Machine Learning model used in the AI system.</li> <li>• Optimization techniques used, like cross validation or boot strapping.</li> <li>• Hyperparameter optimization techniques used.</li> <li>• Hyperparameter values used.</li> <li>• Open-source software, libraries or packages used (if any).</li> <li>• Detailed documentation of data processing and engineering.</li> <li>• The trained parameter values.</li> </ul>
Tier 2	<p>An AI system is considered Transparent for software engineers and tech enthusiasts if it has all the above details plus:</p> <ul style="list-style-type: none"> <li>• Explanations implemented in the AI system show which features played the most important role for a particular prediction.</li> <li>• The AI system has human in the loop.</li> <li>• The technical understanding of human in the loop is known.</li> <li>• The accuracy metrics of the ML model used in the AI system are shared.</li> <li>• The information about the distribution of sensitive variables like gender, race, religion in the data is shared.</li> <li>• The details of different third-party tools used in the AI system are shared, like Google Translate for translations or IBM Watson for conversations.</li> <li>• The detailed documentation of the data used like datasheets and models used like model cards (the assumptions made in the AI development process and disclaimers on the contexts in which this AI system cannot be used) are shared.</li> </ul>
Tier 3	<p>An AI system is considered Transparent for general public if it has all the above details plus:</p>



	<ul style="list-style-type: none"><li>• Explanations are implemented in the AI system in the form of sentences that are easily understandable by general public with minimal technical jargon.</li><li>• A 'human in the loop' is fully responsible for final decision-making.</li><li>• A 'human in the loop' understands the weaknesses of the data used for training ML models.</li><li>• The 'human in the loop' receives a training session on:<ul style="list-style-type: none"><li>o How to use the AI system?</li><li>o The weaknesses of the AI system?</li><li>o When to avoid the AI system?</li></ul></li><li>• User Interface of the AI system takes account of the distribution of sensitive variables in the data and predictions for disadvantaged groups illustrate more information about the limitations of the training data.</li><li>• Weaknesses of different third-party tools used for the AI development are shared, for example Oromo spoken in Ethiopia can't be translated correctly through google translate.</li><li>• Carbon footprint of the energy used in training the models for the AI system is shared.</li></ul>
--	--

# **Chapter 6: Phase 3 - Framework Creation Part 4: Evaluating and Improving the Transparency Index Framework**

## **6.1 Introduction**

The previous three chapters have illustrated how different components of the AI development pipeline were conceptualized for the Transparency Index Framework for application to the education context. This chapter focuses on the evaluation of the Transparency Index Framework that led to the iterative design improvements that are discussed in chapter 7.

In this chapter, I summarize the methodology discussed above for creating the first draft of the framework and then present the methodology for evaluating the framework, including its findings to prepare the final version of the Transparency Index Framework.

The Transparency Index Framework presented here has been developed in three iterative phases of a participatory co-design process. Firstly, an initial Transparency Index Framework was created based on a review of relevant literature and existing frameworks, check lists and tools. Secondly, the Transparency Index Framework was further developed and refined through the participatory design and evaluation of an AI-powered tool that was applied in an educational and training setting. The third phase of the process focussed on a participatory co-revision of the Transparency Index Framework that emerged from phase 2. This third phase engaged a range of different AI in education stakeholders, including educators, ed-tech experts and AI practitioners.

Table 26 shows the mapping of different components of the initial version of TIF to their sources (first two phases of the development): phase 1 as literature review of AI development pipelines and phase 2 as my experience of developing an AI-powered ed-tech for a financial services company. In

phase 3 presented in this chapter, the framework was codesigned and iteratively improved based on AI in education stakeholders' feedback.

Table 26: Mapping of different aspects of the Transparency Index Framework from the literature review and ed-tech tool development process

<b>Components of the Transparency Index Framework</b>	<b>Source</b>
Approach adopted in deploying an AI system	Literature (Chapter 2)
Impact of an AI system	Literature (Chapter 2)
Three Tiers of Users	AI-powered ed-tech tool development process (Chapters 3, 4 and 5)
Data Collection	AI-powered ed-tech tool development process (Chapter 3)
Data Processing	Literature (Chapter 2)
Data Analysis	Literature and AI-powered ed-tech tool development process (Chapters 2 and 3)
Model Selection	Literature (Chapter 2)
Model Training	AI-powered ed-tech tool development process (Chapter 4)
Model Verification	Literature (Chapter 2)
AI Deployment	AI-powered ed-tech tool development process (Chapter 5)
AI Monitoring	Literature (Chapter 2)
AI Improvements	AI-powered ed-tech tool development process (Chapter 5)
Practical Requirements for the Three Tiers of Transparency	AI-powered ed-tech tool development process (Chapters 3, 4 and 5)

## 6.2. Framework Evaluation

The Transparency Index Framework created as discussed in the preceding chapters, was iteratively evaluated in two stages with three groups of stakeholders. The three groups of stakeholders (as shown in figure 21) were identified based on the experience of testing, deploying and making the AI-powered ed-tech tool for the financial services company interpretable, as discussed above. The tech savviness of OM's who were the users and 'humans in the loop' for the tool helped in identifying the measures needed to make the AI's predictions understandable and interpretable for them. These measures were then incorporated in the Transparency Index Framework as shown in table 29.

40 candidates were contacted in two stages to participate in this co-design phase through emails and LinkedIn messaging. These candidates were divided into three groups and short-listed based on their backgrounds, as shown in table 27. 18 candidates responded to this request and participated in interviews. Ten candidates were interviewed in stage 1 and eight candidates were interviewed in stage 2. The questionnaires used for stage 1 and stage 2 interviews are given in appendix 6.

First stage of interviews was used to improve the framework and second stage was used to confirm the effectiveness of the revised framework. Nine candidates requested a copy of the framework and were interested in applying it in their respective contexts straightaway. After stage 1 interviews, some additional details were added to the framework based on the feedback received from educators, ed-tech experts and AI practitioners.

Table 27: Three groups of people interviewed for this research

Group	Description	Number of Candidates
Educators	Teachers, Principals and other Leaders in Schools	9
Ed-tech Experts	People leading the digital strategy initiatives and ed-tech implementations in schools	5
AI Practitioners	AI Practitioners	4

After stage 1 interviews, explorative thematic analysis was conducted with manual coding, and preliminary codes were assigned to the collected data from different interviews (Basit, 2003). Then patterns and themes were identified through the frequency of different codes in interviews from a specific group (educators, ed-tech experts and AI practitioners) and from all the groups combined, to take account of the unique requirements of each group (Belotto, 2018). The Transparency Index Framework was improved based on the findings from stage 1 interviews.

Some of the contributions from stage 1 of interviews are as follows:

- Definitions of different types of biases were added to the framework.
- A table highlighting the assumptions made in grouping different stakeholders of AI in Education (table 28) was added to the framework.
- Language used in the table 25 'Practical requirements for the Three Tiers of Transparency' was made less technical and more understandable for stakeholders with no technical background.

Then stage 2 interviews were conducted as a reliability and validity measure with deductive thematic analysis to confirm the findings from stage 1 interviews. Some of the contributions from stage 2 of interviews are as follows:

- Findings from the stage 1 interviews were confirmed.
- More details were added to the section highlighting the strategic factors that impact the transparency of an AI system and it was added at the beginning of the framework.
- More insights on ethical AI in Education and Transparency Index Framework were gathered.

The findings from stage 1 and stage 2 interviews were incorporated in the final version of the framework presented in chapter 7. The data from the interviews was stored in the cloud for security purposes and ethics approval

for the whole research design was achieved from University College London, Institute of Education.

The interviews with candidates were semi-structured and varied slightly between different groups. This variation was added to ensure that all stakeholders (irrespective of their technical background) understood the implications of ethics and transparency in developing AI-powered ed-tech products. This would enable them to understand the different components of the Transparency Index Framework in a better way and give feedback accordingly. But this also meant that their understanding of AI, transparency and AI ethics can be potentially biased based on my understanding of these concepts and the information I shared with them.

With educators a high-level purpose for the framework was shared during the interviews with only minor details about how the framework development process had evaluated AI tools throughout their development pipeline. Educators were explained what information TIF would show them about a particular AI product and then inquired if they will find such information useful. They were also asked if they have inquired about this information in the past and would they use a framework like this to audit the AI-powered ed-tech products before deploying them in their institutions.

With ed-tech experts, more details of the framework were shared during the interviews compared to educators, and they were asked about the usefulness of this framework as an auditing tool to evaluate ed-tech products. During the interviews, experts were also asked about their opinions on the impact of AI in education, its potential, its impact and any harms it could cause to learners and educators. For the ed-tech experts who were already using AI-powered ed-tech in their schools, they were also asked if they have ever had any conversations on AI ethics or transparency in AI with their ed-tech providers.

With AI practitioners, all the details of the first version of the framework were shared, and their opinions were also incorporated to improve the framework along with educators and ed-tech experts. They were asked about the ethical

considerations in place for their current ed-tech projects and if they were already using any systematic processes for auditing their AI-powered ed-tech tools. They were also asked of any demands they have received from schools for ensuring ethical AI development and how they address these requirements.

## **6.3 Themes**

A set of global themes were shown in interviews by all the stakeholders from different groups, irrespective of their background. These are given in the section below. On the other hand, there were some local themes that appeared across all the stakeholders from a certain group only, with a specific background. For example, these themes emerged from educators with minimal technology background but not from AI practitioners, or some themes appeared in conversations with ed-tech experts but not with AI practitioners.

### **6.3.1 The usefulness of the proposed framework**

Across all the groups, all 18 stakeholders thought that the framework was useful for enhancing their understanding of AI products and where these products can go wrong. Educators stated that the framework could help them get a better understanding of AI-powered ed-tech products and the contexts in which they work best. Some educators who were currently evaluating ed-tech tools were very interested in using the framework as an auditing tool to get a better understanding of these products.

Ed-tech experts also viewed this framework as an auditing tool for evaluating ed-tech products before they are deployed in schools. Most of them relied on GDPR regulations to ensure data privacy but were not aware of any tools or frameworks for AI ethics that are applicable to education.

AI practitioners perceived the framework as a documentation tool to record all the details of the machine learning development pipelines of the

AI products they build. They admitted that some parts of the TIF are more technical than others and would need extra work to make them accessible for all stakeholders of AI in education. For example, one AI practitioner commented on Model Cards in the Transparency Index Framework as “usually more technical summaries, but the visual design (of their representation could be improved)”. AI ethics was a concern for them but their application of AI ethics on products was mostly limited to explanations. From the interviews, it seemed this was the first time they were looking at transparency at a detailed level on every stage of the machine learning development pipeline.

### **6.3.2 Transparency of AI products “a new phenomenon”**

For 17 out of 18 stakeholders that were interviewed for this research, the conversation on transparency regarding ethical considerations in AI products was a relatively new phenomenon. Seven educators involved in the evaluation were using some form of AI-powered ed-tech products in schools, but ethical AI was a relatively new conversation for them. It seemed they were not aware of the adverse consequences of AI going wrong or not working as expected.

All eighteen interviewees, including the AI practitioners building AI products did not perceive transparency in machine learning development pipelines the same way that it was being addressed by the framework proposed in this research. Ed-tech experts seemed to rely on government regulations for protection against AI mishaps. The issues with this approach were that firstly, there are no clear government regulations regarding the bias in AI systems or their malfunctioning, and secondly the government intervention may occur after the damage has been done. Recently, EU has e-published the EU AI Act<sup>16</sup> that aims to address the risks posed by AI applications in real world.

### **6.3.3 Focus on Transparency and Ethics in AI products**

---

<sup>16</sup> <https://artificialintelligenceact.eu/>



Transparency in AI was a relatively new phenomenon for all the stakeholders. The concept of making their AI implementations transparent throughout the machine learning development pipeline and sharing it with end-users was new to AI practitioners.

Three practitioners made an interesting point regarding transparency in AI stating that they do not put in a lot of effort in making their AI products transparent because their clients (education institutions like schools) do not ask for it. This theme also emerged in my interviews with educators and ed-tech experts in schools: they were mostly not aware of the importance of transparency in the AI products they use in schools. Although one practitioner confirmed that in the past, they have received questions from teachers when teachers noticed something strange like why the tool is making this particular decision for this student. Such questions from teachers highlight the importance of transparency for AI products. In certain contexts, as in this case, lack of transparency can lead to confusion or unpleasant experience for teachers in the classroom.

All four AI practitioners that were interviewed in this research claimed that they try to make their machine learning models explainable. They focus on post-hoc explainability through tools like Lime (Ribeiro et al, 2016). But, they have never received particular requests from educational institutions on adding explainability to their AI products or sharing the details of their development process.

Ed-tech experts identified the importance of ed-tech and AI in ed-tech to enhance learning outcomes for students. In their opinion, the claim often made by AI-powered ed-tech products that they reduce teacher workload is not backed up by evidence. They claimed that AI in ed-tech is over-rated and not as impactful as some companies claim.

When asked about their concerns regarding AI going wrong and negatively impacting learning outcomes based on false information or wrong predictions, two contrasting themes emerged from ed-tech experts:

- They relied on regulations like GDPR to take care of any such mishaps. Ed-tech companies operating within EU need to make their products compliant with GDPR. They thought that GDPR would also take care of any ethical, accountability and transparency issues within AI along with data privacy and storage concerns.
- Another theme that emerged in the discussions with ed-tech experts focused on AI hype in ed-tech products. They thought that ed-tech companies claiming to use AI in their products exaggerated the benefits of AI. According to them, any major breakthroughs in AI would require huge amounts of investment. These ed-tech experts were also comparatively more aware of the importance of ethical AI in education. They claimed that the unintended consequences of AI in education are not as well documented as in other sectors. This poses more danger as most educators, and sometimes even ed-tech companies are not very well aware of where AI can go wrong in the context of education and the impact this can have on learners. It is important to note that the ed-tech experts with these beliefs thought that there might have been mishaps in AI in education, but they are not very well documented.

For educators, transparency in AI was a new phenomenon. They were excited about trying new ed-tech and AI products in their schools and evaluating their impact on learning outcomes, but mostly did not seem concerned about AI's negative consequences. Most of them were not tech savvy, but they understood the purpose of this research's framework and how it could be useful in their contexts.

For all nine educators interviewed, this was the first time they were having conversations on ethical AI and what kind of documentation or precautionary measures to expect from ed-tech companies applying AI in education. It seemed they have never had such conversations from their ed-tech providers earlier. A theme that was also confirmed in an interview by the AI practitioner leading a data science team at one of the biggest ed-tech companies in the world.

#### **6.3.4 Feedback and Recommendations to Improve the Framework**

AI practitioners were the most active group in terms of providing feedback to improve the framework. This can partially be because most of the literature I draw from to create the initial version of the framework comes from industrial and technical perspectives as shown in chapter 2 and table 26. AI practitioners recommended to add a brief summary of each type of bias that can exist in a machine learning development pipeline. They also advised adding another clause regarding open-sourcing or publicly sharing the development code that was used to build the product. This would enable any watchdog or AI auditing group to replicate the results of that AI product and identify any gaps with sample data.

AI practitioners showed the most interest in the details of the framework, especially the data processing stage. AI practitioners requested to view a copy of this framework and showed interest in using this research in their projects when it is published as well. They also recommended to highlight the assumptions made in categorizing the different stakeholders of AI in Education into three tiers.

Most educators wanted to have follow up conversations on the framework. They wanted to discuss it with their colleagues and incorporate some of the questions in teacher training sessions to enhance their understanding of AI products before they are used in classrooms. From the conversations, it seemed that the framework helped educators

in identifying a gap in their current auditing process for AI-powered ed-tech products before they are deployed in schools.

Some educators also asked to view a copy of the framework to give feedback. One educator pointed out that the Transparency for Tier 3 users (general public: teachers, students and parents) as depicted in the Transparency Index Framework “is too esoteric and more linked to their perceived outcomes”.

Based on the educators’ feedback, a separate version of the framework can be prepared specifically for schools. This version of the framework will have the definitions of all the technical terms. It can also explicitly mention how the answers to certain questions in the framework can indicate an AI tool’s effectiveness in enhancing educational outcomes. For example, if no usability tests have been done with sample schools before selling an AI-powered Learning Management System at scale, it can be considered a red flag for the schools to buy such an AI tool.

Another educator requested that the language used in the framework needs to be “explicated/simplified for ease of access” by non-technical audience like educators and school leaders. They particularly requested to add the definitions of different types of biases in the framework.

## **6.4 Conclusion**

The interviews with different AI in education stakeholders discussed in this chapter provide invaluable insights and feedback regarding their opinions on AI in education, the importance of ethical AI, conceptualization of transparency for AI-powered products in educational contexts and changes to the Transparency Index Framework that would make it more accessible, practical and understandable for educators, ed-tech expert and AI practitioners.

The final version of the framework prepared after thorough consultation with these stakeholders is presented in the next chapter.

# Chapter 7: Phase 3 - The Revised Transparency Index Framework

## 7.1 Introduction

This Chapter 7 presents the Transparency Index Framework that has been revised in the light of the evaluation conducted with various AI in education stakeholder and discussed in Chapter 6.

The revised Transparency Index Framework can be adopted as a continuous approach where transparency is not seen as an instantaneous decision or as dependent on the usage of a particular set of tools only. It is adopted as a continuous process, integrated into the design usage scenarios of an AI tool. There are several factors that influence the type of transparency that should be or can be induced in an AI system.

## 7.2 Framework Description

The Transparency Index Framework has three components that address different aspects of an AI-powered tool. These components are as follows:

1. Checklist: provides a set of questions and guidelines for AI practitioners and ed-tech companies to incorporate transparency in strategic decision-making, machine learning modelling and deployment stages of the AI development process. This part of the framework was derived from phase 1 (literature review) and phase 2 (development of an AI-powered ed-tech tool) of this research study.
2. Categorization of stakeholders: categorizes the different stakeholders of AI in education into three groups based on the tech savviness of the stakeholders and the resources required to make an AI-powered ed-tech tool transparent for them. These three categories of stakeholders were identified during phase 2 of this research. In the AI development process of an ed-tech tool, there were three main categories of stakeholders from the financial services company who were involved: firstly, office managers and domain experts who would use the tool and with whom the knowledge elicitation process of financial services and trading was conducted.

Secondly, the software engineers and a tech team who were responsible for the technical development of the AI tool. Thirdly, the data team (including an AI practitioner) to whom all the knowledge was transferred and the minimum viable product of the tool was handed over for maintenance and iterative improvements later.

3. Requirements of transparency for different tiers of stakeholders: highlights a set of guidelines that can be followed to make an AI-powered ed-tech tool transparent for different tiers of stakeholders. This section of the Transparency Index Framework was derived from the experience of developing an AI-powered ed-tech tool. It was validated and improved based on the interviews with different AI in education stakeholders in phase 3 of this research.

### **7.3 Transparency Index Framework**

The process through which Transparency is ensured in an AI system or the extent to which Transparency is needed for an AI system is determined by a number of strategic decisions that should be taken into account before starting the data collection process for an AI system. These are as follows:

#### **Strategic Decisions**

1. What approach will you adopt in deploying the AI system to production:
  - 1.1. **Human in the Loop:** Final decision-making authority is kept with the humans, AI system's role is to enable more informed decision making.
  - 1.2. **Human on the Loop:** Human plays a supervisory role to evaluate the decisions made by an AI system, before and they are implemented in real-world.
  - 1.3. **Human out of the Loop:** AI system makes the decision with no human involvement.
2. What kind of impact will this AI system have on its users:

- 2.1. **Direct Impact:** where a user's personal life is affected by a decision and in some cases user has no choice other than compliance, like recruiting a candidate, giving a loan or deciding recidivism.
  - 2.2. **Indirect Impact:** where a user's day to day life is not affected by an AI system's decision or user has the choice to decide against the AI's decision, like spam filtering in emails, or recommendations on an e-commerce store.
3. What is the tech savviness of the individuals who would be (directly or indirectly) impacted by this AI system:
- 3.1. **Tier 1: Researchers and AI practitioners:** they have a thorough understanding of the techniques that are needed for the development of a machine learning powered AI system.
  - 3.2. **Tier 2: Software engineers and tech enthusiasts (ed-tech experts and tech enthusiasts):** they have some understanding of the techniques that are needed for the development of an AI system.
  - 3.3. **Tier 3: General public (teachers, students and parents):** they do not have any understanding of the techniques that are needed for the development of an AI system.

## 1. Data Transparency

### 1.1. Data Collection

- 1.1.1. How was the data collected?
- 1.1.2. What were its sources?
- 1.1.3. Was Assumption Testing carried out: What assumptions were made regarding the data collection?
  - 1.1.3.1. How many of these assumptions were tested and verified?
- 1.1.4. Was informed consent taken from all individuals?
- 1.1.5. What data on sensitive variables like gender, nationality, ethnicity and religion etc is collected?
- 1.1.6. How is your data labelled:
  - 1.1.6.1. By ground-truths.
  - 1.1.6.2. By human labels.



1.1.7. From 1 to 10, how do you rate the involvement of domain experts in data collection? For example, if you are building an AI-powered product for teachers, did you consult any experienced teachers regarding what you are building, and the kind of data you are collecting for it.

## **1.2. Data Processing**

1.2.1. How is data stored and ensured that it is secure?

1.2.2. How was data normalized?

1.2.3. What techniques and tools were used to process the data?

1.2.3.1. Were the strengths and weaknesses of these techniques explored?

1.2.4. How was the sensitive variables data processed?

## **1.3. Data Analysis**

1.3.1. What techniques and tools were used to analyze the data?

1.3.2. Was Exploratory data analysis done?

1.3.2.1. Was Exploratory data analysis shared and confirmed with domain experts?

1.3.3. Was statistical data analysis done?

1.3.3.1. Was statistical data analysis shared and confirmed with domain experts?

1.3.4. Were correlations between different features identified and confirmed by domain experts to evaluate any assumptions made?

1.3.5. What techniques were used to identify, mitigate and share the limitations of data with stakeholders?

1.3.6. What types of biases were identified in the data:

1.3.6.1. Historical Bias: This bias exists in society and is reflected in the data even if there are no errors in the data collection and processing stages.

1.3.6.2. Representation Bias: This bias occurs when sample data used to build the AI is not truly representative of real world.

1.3.6.3. Measurement Bias: This bias occurs while choosing, collecting and computing features in the data to measure a certain outcome.

- 1.3.6.4. Aggregation Bias: This bias occurs when one size fits all AI approach is used for all the groups in the data.
- 1.3.6.5. Evaluation Bias: This occurs when the test data of an algorithm does not represent the target population.
- 1.3.6.6. Deployment Bias: This occurs when there is a mismatch between the problem that an AI tool is built to solve and what it is actually used for in the real-world.
- 1.3.7. What steps were taken to mitigate the above biases in the data?
- 1.3.8. Are domain experts informed of the measures taken to mitigate these biases?
- 1.3.9. Are domain experts fully briefed on the potential impact of each type of bias on AI system's predictions?
- 1.3.10. Was Datasheet prepared for the data processing stage?

## **2. Algorithmic Transparency**

### **2.1. Model Selection**

- 2.1.1. Which ML model was used for predictions?
- 2.1.2. Why was this particular model chosen?
- 2.1.3. Was any experimentation done with different models? Was a Models Evaluation Report prepared?
- 2.1.4. What are some common strengths and weaknesses of this model?
- 2.1.5. In choosing the model were Transparency capabilities of the model taken into consideration?
- 2.1.6. Is the model doing regression or classification?
- 2.1.7. Is the model using any Explainable Artificial Intelligence (XAI) tools or providing explanations of the predictions:
  - 2.1.7.1. If yes, which XAI tools are being used?
    - 2.1.7.1.1. What are the strengths and weaknesses of these tools?
    - 2.1.7.1.2. Was human in the loop trained regarding the limitations of these explanations?
    - 2.1.7.1.3. Were any measures taken to address these limitations in autonomous AI systems?

### **2.2. Model Training**

- 2.2.1. Which tools (like Python libraries) were used for training the models?
- 2.2.2. What hyperparameters were used for training the models?
  - 2.2.2.1. Were these hyperparameters optimized?
    - 2.2.2.1.1. If yes, what techniques were used for hyperparameter optimization?
- 2.2.3. What was the percentage of training and test set?
- 2.2.4. Was the distribution of features in training and test set similar?

### **2.3. Model Verification**

- 2.3.1. How was the machine learning model audited. For example, what were the results of using counterfactuals etc?
- 2.3.2. Have you tested your model on a subset of sensitive variables data?
- 2.3.3. Have you prepared a disclaimer document highlighting the exact contexts in which your model can be used?
- 2.3.4. Was a Model Card prepared for your models deployed in real-world?

## **3. Implementation Transparency**

### **3.1. AI Deployment**

- 3.1.1. Have you tested the MVP of this AI system with potential real-world users?
  - 3.1.1.1. Were the domain experts or users satisfied with the tool's performance?
  - 3.1.1.2. Will you share the details of this MVP testing with prospective clients?
- 3.1.2. Is there some form of visual signaling to indicate that a particular aspect of this AI system is a work in progress, or is not perfect or have certain biases against these particular groups?
- 3.1.3. What is the carbon footprint of training the ML models being used in this AI system? For example, some cloud providers provide information on environmental impact of training a particular ML model.

### **3.2. AI Monitoring**

- 3.2.1. How will you be monitoring this AI system in production?
- 3.2.2. What security and privacy measures were taken when deploying the AI system?
- 3.2.3. Have you prepared a Models Validation Report to document the tool's performance in real-world with focus groups or the first few users?
  - 3.2.3.1. Were the results up to expectation?
    - 3.2.3.1.1. If not, what changes were made in the AI system?
    - 3.2.3.1.2. Were steps 3.2.3. onwards repeated unless the AI system reached expected results?

### 3.3. AI Improvements

- 3.3.1. How often are you planning on pushing the improved model to production?
- 3.3.2. Have you identified the lower limit below which the AI system needs attention or human intervention?
- 3.3.3. Have you identified the lower limit below which the AI system should stop working?
- 3.3.4. Have you completed registration or acquired endorsements (like completing conformity assessments) from regulators or other third parties, like registration on public EU database for high-risk AI systems?
- 3.3.5. Have you prepared the Factsheet for your AI tool?

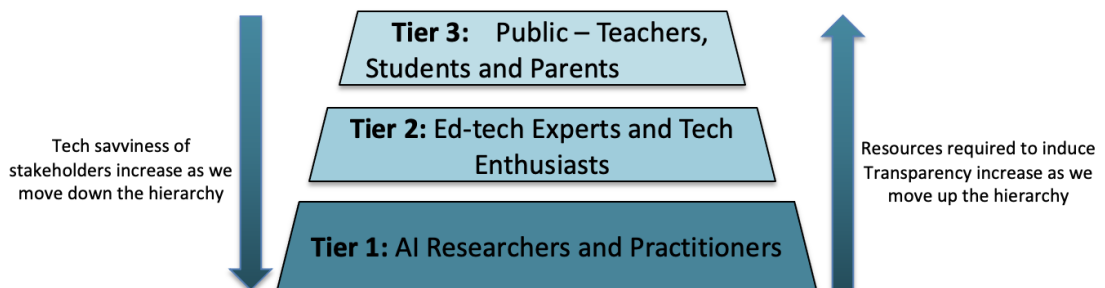


Figure 21 (above): Three Tiers of Transparency for AI in Education

Table 28 (below): Assumptions / Characteristics of the Three Tiers of Transparency

Tier 1	Tier 2	Tier 3
Have extensive technical knowledge of how AI systems work	Have basic knowledge of AI applications in real-world like NLP, image recognition	Tech knowledge is limited to the use of social media or other personal apps like Uber etc
Can reproduce AI research from research papers	Have know-how of AI's applications in education and big players in this space	Not tech savvy and no technical background
	Have extensive experience of using and deploying ed-tech in schools	Not aware of the limitations of AI or technology in general
		Need thorough training before using tech products

Figure 21 maps the different types of stakeholders in education in terms of their technical background to the resources required to make an AI system transparent for that particular group. It shows how the requirements for transparency in AI products vary with the background of stakeholders. To make an AI system transparent for stakeholders with no technical background, like educators, AI companies need to invest significant resources and time. This part of the Transparency Index Framework was devised in phase 2 during which various measures were taken to make the tool transparent for Office Managers who were acting as 'humans in the loop'.

Table 29 in the Transparency Index Framework shows the measures that the ed-tech companies need to take for making their AI development processes transparent for different stakeholders of their product. It shows how the measures needed for Transparent AI development increase as we move from tier 1 to tier 3 stakeholders. This table was derived from my experience of developing an AI-powered ed-tech tool, and then evaluated through interviews with different stakeholders.

Table 29: Practical requirements for the Three Tiers of Transparency

Tiers	Description
Tier 1	An AI system is considered Transparent for AI researchers and practitioners if they know or have access to the:

	<ul style="list-style-type: none"> <li>• The Machine Learning model used in the AI system.</li> <li>• Optimization techniques used, like cross validation or bootstrapping.</li> <li>• Hyperparameter optimization techniques used.</li> <li>• Hyperparameter values used.</li> <li>• Open-source software, libraries or packages used (if any).</li> <li>• Detailed documentation of the data processing and engineering.</li> <li>• Trained parameters values.</li> <li>• Open-source code (or a relevant part of it) of the AI system.</li> </ul>
Tier 2	<p>An AI system is considered Transparent for ed-tech experts and enthusiasts if it has all the above details plus:</p> <ul style="list-style-type: none"> <li>• Explanations implemented in the AI system show which factors played the most important role for a particular prediction.</li> <li>• The AI system has a human in the loop who understands the explanations.</li> <li>• The technical understanding of human in the loop is known.</li> <li>• The accuracy metrics of the ML model used in the AI system are shared.</li> <li>• The information about the distribution of sensitive variables like gender, race and religion in the data is shared.</li> <li>• The details of different third-party tools used in the AI system are shared, like Google Translate for translations or IBM Watson for conversations.</li> <li>• The detailed documentation of the data used, such as datasheets, and models used, such as model cards (including the assumptions made in the AI development process and disclaimers on the contexts in which this AI system cannot be used) are shared.</li> </ul>
Tier 3	<p>An AI system is considered Transparent for educators and parents if it has all the above details plus:</p> <ul style="list-style-type: none"> <li>• The AI system has been thoroughly tested with sample users and their findings have been incorporated in the product development and training.</li> </ul>

	<ul style="list-style-type: none"> <li>• AI explanations are implemented in the AI system in the form of sentences with minimal technical jargon and are easily understandable by general public.</li> <li>• A ‘human in the loop’ is fully responsible for final decision-making.</li> <li>• A ‘human in the loop’ understands the weaknesses of the data used for training ML models.</li> <li>• A ‘human in the loop’ can explain the workings of an AI system to users.</li> <li>• A ‘human in the loop’ has a thorough understanding of when to rely on an AI system and when to avoid it.</li> <li>• The ‘human in the loop’ receives a training session on: <ul style="list-style-type: none"> <li>o How to use the AI system.</li> <li>o The weaknesses of the AI system.</li> <li>o The contexts in which to avoid using the AI system.</li> </ul> </li> <li>• User Interface of the AI system takes account of the distribution of sensitive variables in the data. Predictions for under-represented groups illustrate more information about what the training data lacks about such groups that might skew the results for a particular individual.</li> <li>• Weaknesses of different third-party tools used in the AI system are shared if relevant, for example Oromo spoken in Ethiopia can’t be translated correctly through google translate.</li> <li>• Carbon footprint of the energy used in training the models for the AI system is shared.</li> </ul>
--	--

## 7.4 Discussion: Transparency in the context of AI in Education

Building impactful and effective AI for education is hard (Kay, 2012). If we take a simple AI in education use-case to facilitate learners or enhance their understanding of a particular concept like algebra (Chien et al, 2008; Beal, 2013) or fractions (Beal et al, 2010) or grammar (Alhabash et al, 2016), these are difficult problems to solve for AI at scale. Teaching students can be

challenging for humans too and not all human teachers are necessarily great at explaining these concepts to all students. A great teacher for one learner might have no impact on the learning outcomes of another student because every learner is unique and a particular teacher's pedagogical style might be more suitable for some learners than others. These unique teaching styles and pedagogical methods along with varying learning needs of different students makes it very important and challenging for AI systems like Intelligent Tutoring Systems (ITS) to provide contextualized learning to students. For AI systems deployed in educational contexts, it's very important to thoroughly document the data they are trained on because this training data plays a significant role in determining the kind of contexts that AI system would work in as expected. The data transparency section of the Transparency Index Framework aims to fill this gap by providing a detailed guideline on documenting the data processing stage of the AI development process.

For AI deployments in education, the Transparency Index Framework proposed in this thesis encompasses different aspects of the AI development process that AI practitioners need to document. For example, for ML modelling, the details of the models used for decision-making can be documented and shared. To take this one step further, the ed-tech companies developing AI-powered products can also choose to publicly share the code that their AI practitioners and/or data scientists write. This can enable reproducible research and is a contribution to the tech community working on AIED.

Some might argue that steps like making the code of AI implementations public through GitHub or other tools is not very helpful for general public or Tier 3 users mentioned in the framework like educators or ed-tech experts who are not tech savvy and at times not interested in the technical details of the development process. But, such steps may help the tier 3 users indirectly. This has both a push and a pull factor for AI practitioners working on ed-tech products as they know their work (code) will be viewable by public in future which reinstates the need to work towards public good (Elster; 1998;



Chambers, 2004; Chambers 2005;Naurin, 2007). It also means that practitioners know they can be held accountable for their work which can lead to more robust AI development. Hence, these requirements were added in the TIF.

Other researchers have also argued against complete transparency like Zarski (2016), Lepri et al (2017), De Laat (2018) and Carabantes (2019). Complete transparency like making the code of an AI tool public can hinder innovation as companies will potentially be sharing the intellectual property of high commercial value that makes AI work in certain contexts and provides them competitive edge over others. In such scenarios it can be argued that the ed-tech company developing an AI product can implement the different components of Transparency Index Framework for all three stages of the AI tool development process but can avoid sharing all such information with end-users or third parties. They induce transparency for internal use. Hence, this research addresses the criticisms of complete transparency by creating a distinction between transparency for the ed-tech companies developing the tools and transparency for external stakeholders like end-users of the ed-tech.

Another problem with complete transparency, especially in education where most stakeholders are in tier 3 of the TIF, is that it can potentially lead to information overload for stakeholders (Eppler and Mengis, 2004) or transparency paradox (Richards and King, 2013). Sharing everything with the stakeholders can be counterproductive, potentially lead to confusion for stakeholders and make it more difficult for them to find the relevant information (Stohl et al, 2016). AI-powered ed-tech companies for teachers face this risk as teachers in a classroom setting can be easily confused by too much information on their dashboards (Bull et al, 2013; Greller and Drechsler, 2012). Luckin *et al* (2006) have illustrated the importance of human centered design in developing educational systems that are fit for use. They highlight the importance of iterative improvements in building such educational systems. Considering the risks involved, this particularly holds true for any kind of AI system in an educational setting. AI-powered ed-tech tools need to be thoroughly tested with target users as a part of the participatory design

approach before they are deployed in the real world at scale. It is added in the 'Implementation Transparency' section of the Transparency Index above.

At times the users of an AI-powered ed-tech may not even know what information they need. What is useful for them, what kind of impact lack of transparency can have on them or what is too much transparency for them that leads to cognitive overload (Bogina et al, 2021). This is especially the case for tier 3 users who are not tech experts and do not know exactly what kind of information from the entire AI tool's development pipeline will be useful for them. This is also shown from this study, that educators who are also tier 3 users (according to the framework) have mostly never had conversations on ethical AI and/or transparency in AI before. This was also confirmed from AI practitioners who stated that they have never received any requests on transparency or concerns regarding ethical aspects of machine learning powered AI in product development from their clients (educators and ed-tech experts). In such industries, Transparency Index Framework for the ed-tech companies developing AI tools is more important to mitigate the risks of AI as there seem to be no external checks in place by other stakeholders.

A counter point to the above argument is that even if sharing the development details of an AI tool leads to cognitive overload, this does not mean that ed-tech companies should stop making such information public at all. End-users of an AI tool do not necessarily need to know or understand every detail of an AI implementation, but this belief in AI practitioners that they need to share every decision and assumption made during the tool's development can act as a strong precautionary measure for them to double-check these decisions, leading to more robust development processes. These checks and balances can also prevent mistakes which lead to unexpected and controversial results and can be harmful to the ed-tech company's image.

The criticisms on transparency are mostly directed at information that is shared with end-users. If the focus is on the question of 'transparency for whom' and the transparency measures to be taken by ed-tech companies when developing AI are treated separately from the information that they need

to share with various stakeholders of education, then it can be noticed that the criticism is mostly directed at the information shared with end-users, not the measures to be taken by ed-tech companies for internal transparency. For example, the autopilots working in cars are powered by state-of-the-art image recognition algorithms trained on vast amounts of data (Hirz and Walzel; 2018). When drivers are using the autopilots, they do not necessarily want to know the detail of how AI is making every decision or identifying different road signals etc. They need to know when not to trust the AI system, for example during heavy rainfall etc. Similarly a teacher using AI-powered learning analytics dashboard with real-time data does not need to know how it is processing the data or making every prediction, but when not to trust the system, for example when the tool is making predictions for a student with under-represented ethnicity, demography or age in the training data. But, if the companies developing such software do not feel the need to share the details of their AI system with end-users, it does not imply that they should ignore the transparency considerations while developing that AI system. As shown in this research, these transparency considerations also lead to robust and well-documented AI systems that are very beneficial for the ed-tech companies developing these tools. For ed-tech companies, such measures can save time, resources and man-power required to debug the AI systems, improve their performance and make them suitable for different educational contexts.

Ed-tech companies need to allocate resources to develop ethical and safe AI for their users. If a particular set of information regarding an AI tool is not to be shared with end users because it may lead to cognitive overload or reveal company's secret sauce of building effective AI, it does not mean that the company should stop implementing the transparent development measures shown in the Transparency Index Framework. Documentation of the decisions taken, and assumptions made during the AI tool development process can be very useful for the company itself (Madaio et al, 2020) as shown above.

Cognitive overload in the context of Transparency Index Framework can be caused by sharing all the information and documentation of the data

processing, ML modelling and deployment stages with the users of an AI system (Kirsh, 2000) such as teachers, head teachers and learners who are not tech savvy and sometimes not interested in these details. But, there are a number of ways in which cognitive overload for such users can be avoided without compromising on the principles of transparency. For example, despite the documentation of the entire AI tool's development pipeline, only relevant information can be shared with the stakeholders. If this information is too much, it can be shared over a period of time or made available to stakeholders and left it at their discretion to access it, as and when needed. Cukurova *et al* (2019) have presented a framework for evidence-informed educational technology where ed-tech companies work closely with researchers and educators to ensure the efficacy of the products they build. For safety in AI systems, a participatory design methodology where ed-tech companies closely work with the prospective users of their AI offering to understand their needs is necessary (Luckin et al, 2011). This is where the 'Implementation Transparency' section of the AI Transparency Index plays a crucial role.

There are many ways in which transparency informs and overlaps with other dimensions of ethical AI. Figure 22 shows how transparency as presented in the Transparency Index Framework can overlap with explainability, fairness, accountability, interpretability and safety of AI systems.

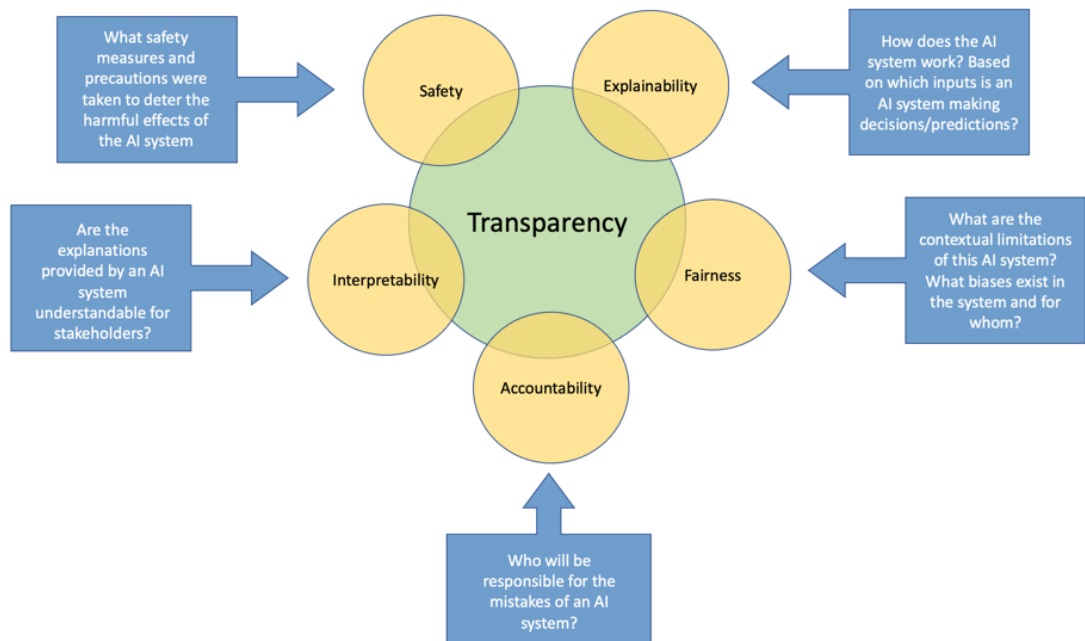


Figure 22: Transparency in relation to the other dimensions of ethical AI as depicted in the Transparency Index Framework presented in this research

Explainability and Interpretability of AI systems are at times used interchangeably (Linardatos et al, 2021). They seem to be a necessary prerequisite for transparent AI systems as well. Ed-tech companies do not need to share all the details of their AI tools with end-users as discussed in this section above. But, the information they do share needs to be understandable by the users of the AI system. Considering the importance of explainability of AI systems in educational context, it was explicitly added as a requirement in the Algorithmic Transparency stage of the Transparency Index Framework. To ensure the understandability of these explanations by teachers or learners, usability testing of AI systems was added in the 'Implementation Transparency' stage.

Any AI system developed in the lab and/or deployed in the real world only works as expected, or fairly, in certain contexts and based on certain assumptions. Economists use the terms 'ceteris paribus' to indicate how economic laws work in certain conditions when all external factors are in perfect equilibrium (Schiffer, 1991). The same can be applied to AI systems in the real-world as well. To ensure fairness in AI systems, transparency enables the companies developing AI-powered ed-tech products to highlight the

assumptions in which they expect the tool to perform fairly or at its optimum level. For example, a learning analytics dashboard that visualizes the learning progress of different students in classroom may be designed and tested in classrooms with one teacher and less than twenty students in private English schools with a vast majority of white English students. It may not work as effectively when deployed in a classroom with two teachers (an additional assistant teacher) and more than thirty students who are a mix of different ethnic backgrounds. Therefore, thorough testing of the MVP of an AI system with prospective users in real-world was added as a requirement in the 'Implementation Transparency' stage of the AI Transparency Index. Hence, transparency seems to be an integral part of fairness in AI systems in educational contexts.

According to some researchers, transparency and accountability are closely related (Hood, 2010; Matthias, 2004). From the viewpoint of ed-tech companies, there are strong reasons why accountability leads to more transparent AI development processes and vice versa. Accountability in terms of transparency for ed-tech companies is a two-edged sword. On one hand, ed-tech companies might want to make their development processes transparent and share all the details (pros and cons) with educational institutions to avoid taking any responsibility of mishaps. But, on the other hand they may want to avoid sharing any details with educational institutions because it makes their product difficult to sell and they might be held accountable for the weaknesses or misuse of their products. Transparency as depicted in the TIF for AI-powered ed-tech can help companies document and test the assumptions on which they build their AI systems and hold the relevant departments or individuals accountable when tools perform unexpectedly.

Within different dimensions of ethical AI, safety is of utmost importance and cannot be compromised. It quite often gets ignored in the race to be the first one to launch an AI-powered product (Amodei et al, 2016). In education, safety is even more crucial because of high stakes and because mishaps in AI-powered ed-tech can go unnoticed unless a teacher or student raises it

with the relevant authorities in their school, as mentioned by one of the ed-tech experts during the interviews. Safety of AI systems, like transparency, is not dependent on a particular tool but needs to be ingrained in the whole design and development process. Transparency Index Framework presented in this research can lead to more robust AI systems because it encourages the ed-tech providers to document the details of their data processing, machine learning modelling, user testing and deployment stages. This documentation enables AI practitioners to test their assumptions, justify their decisions and adopt a more cautious approach when selecting different technical tools, libraries and third-party services in building their products. Hence, leading to safer AI systems.

The Implementation Transparency stage of the Transparency Index Framework plays an important role in identifying and documenting the target users of an AI-powered ed-tech and the ideal contexts for it to perform as expected. This helps in highlighting the diversity and inclusion limitations of AI systems and can be considered the first step towards more accessible AI in education. For example, TIF can assist in identifying an AI-powered student drop-out prediction tool that has been trained on student data from three universities in south England with mostly white students. Its applications in Asian universities with different ethnicities and academic culture may not produce similar results.

The inclusion and diversity limitations of AI systems are not necessarily intentional, but mostly based on the historical biases that exist in the society (Roselli et al, 2019; Fuchs, 2018). In AI systems in educational contexts, we face the risk of aggravating these biases if the diversity and inclusion limitations are not explicitly addressed. Stathoulopoulos and Mateos-Garcia (2019) have proposed gender diversity in the teams that build AI systems. It can be extended to the diversity of race, ethnicity, religion, nationality and culture in the design and development teams of AI-powered ed-tech products to take account of the diverse viewpoints and belief systems (Kolbjornsrud et al, 2016).

Some researchers argue that business leaders and AI practitioners should have ethical considerations in place when conceptualizing an AI tool to empower their ed-tech (Eitel-Porter, 2021). Ethical AI is a very broad term, with many different connotations and dimensions that each have their own complexity, as discussed in chapter two. It can be challenging for business leaders and even AI practitioners to translate the broad concept of ethical AI into implementable actions that would ensure well-being of all the stakeholders involved. Transparency, as conceptualized in different stages of the Transparency Index Framework, narrows down the broad horizons of ethical AI into more actionable steps that (if implemented and considered appropriately) can also enable other ethical AI dimensions like explainability, interpretability, accountability, fairness and safety into their product development processes.

Theoretically, transparency in AI does not replace ethical AI, it is its subset. It may not be a sufficient condition for ethical AI according to some experts, but it is a necessary condition for ethical AI in education (The Institute for Ethical AI in Education, 2021). If there is one dimension of ethical AI that AI practitioners need to choose to focus on, it can potentially be transparency in AI for several reasons. Firstly, transparency covers the entire AI development process from initial designs to deployment in real-world. It co-exists with the AI tool as shown in the Transparency Index Framework. Secondly, transparency can ensure (as seen from the Transparency Index Framework) that the other ethical AI dimensions like explainability, safety and fairness are being addressed as well. Thirdly, transparency facilitates and benefits both internal and external stakeholders of an AI-powered ed-tech: a) ed-tech companies through a thorough documentation of the tool and robust development processes, and b), AI system users through a better understanding of how the AI system works.

A lot of criticism of AI systems in education like ITS has been due to their focus on pedagogy rather than learner agency (Herold, 2017; Watters, 2015; Watters, 2017; Wilson and Scott, 2017). Ed-tech companies can build very impressive tools based on the latest AI techniques but often do not evaluate



the effectiveness of their impact. Their products may not be evidence informed. Therefore, the deployment and iterations stages of the AI development process in the Transparency Index Framework can play a very important role in determining the effectiveness of an AI-powered ed-tech product. It gives learners (if they are the intended users of AI) a pivotal role in designing effective AI-powered products. Validating the Minimum Viable Product (MVP) of the AI tool with potential real-world users can help in identification, mitigation and (at times) removal of any unintended effects in an AI system before it is deployed in real-world at scale.

The tier 3 users' viewpoints on AI can be impacted by sometimes the hype created by media. It can be misleading to conclude that if AI can master the game of Go (Silver et al, 2016) or solve rubric cube (OpenAI, 2019), it can do anything better than humans. This can be dangerous for tier 3 users in the Transparency Index Framework who are not mostly aware of the mishaps of AI in real world, its limitation and the impact AI's mistakes can have on educators and learners. For example, it is intuitive for a teacher who has recently read about AI beating humans in the most complex board game (Go) to also be better in helping students, diagnosing learning gaps or making predictions about learners' performance. AI hype can easily lead to over reliance on AI systems' decisions or automation bias. Therefore, the training of the 'human in the loop' or people using the AI system was added in the Transparency Index Framework. The purpose of this training should also be to empower the educators to trust their own good judgement over AI's decision whenever there is a conflict.

I have made the distinction of transparency between the ed-tech companies developing AI tools for themselves and transparency for the users of their AI products. For the companies building AI-powered ed-tech products, transparency of every component of each cannon in [figure 7](#) is very important. AI development process is complex and time-consuming where tens and at times evens hundreds of domain experts, software engineers and AI practitioners may work on the same product together. This creates team dynamics that are difficult to maintain efficiently. Hence, every team or

individual's contribution needs to be documented and recorded for future reference. This also makes the debugging and maintenance of the AI system a lot more efficient and cost-effective. Therefore, it is very important and beneficial for ed-tech companies to implement the processes mentioned in the Transparency Index Framework that make their AI development process transparent, at least for themselves.

Companies who want to be transparent with users about their AI-powered ed-tech offering may find a number of ways of sharing their information with different stakeholders. It can also be a part of their sales strategy as publicly shared information can add to company's credibility in terms of the confidence they have in their product. For example, publishing a part of the software code of their product on Github so others can benefit from it, test it or use it to audit their products can give a positive message of company's AI offerings. Similarly, companies may choose to publish the impact or effectiveness of their products through research papers in conferences or research journals to get accreditation from academia. For example, ASSISTments<sup>17</sup> has used this approach to validate the effectiveness of their mathematical interventions (Heffernan and Heffernan, 2014; Selent et al, 2016; Cirella, 2021).

Regarding the transparency for end-users, ed-tech companies can choose to be transparent about specific pieces of information with their end-users. This may depend on a number of factors such as the regulatory obligations in the countries in which the company operates. For example, under GDPR in Europe, companies collecting or using individual's data in any form have to share certain information with that individual. This may cover some aspects of the data processing stage, but not all. Intellectual property of the company and the role it plays in that company's valuation may play an important role in determining the kind of information about an AI system that the company shared with end-users or other stakeholders. As AI is a rapidly developing field, a particular ed-tech company may have developed a secret sauce that gives their product a unique competitive advantage over others. It should not

---

<sup>17</sup> <https://new.assistments.org/>

prevent them from applying transparency on their AI development process as conceptualized in the Transparency Index Framework.

Competition that the company faces or its long-term vision may also impact the company's strategy in being transparent about the development process of their AI product. In a saturated domain like Intelligent Tutoring System for Algebra or AI-powered English grammar tool, companies might be wary of sharing too much information due to competition. On the other hand, if a company envisions to be a not-for-profit that aims for bigger social good, they may want to make as much information public as possible to benefit others in society as well. Cognitive overload of end-user is another important reason why ed-tech companies may limit the kind of information they share with stakeholders as discussed above. This is especially true for AI-powered learning analytics that empower teachers with real-time data during classroom lessons. In such situations it may be very easy for teachers to get overwhelmed with too much information. TIF aims to address these concerns by covering different aspects of the AI development process in a manner that makes their documentation useful for the ed-tech company's internal use. The extent to which this information is made public is dependent on several factors and should not necessarily determine whether the company should implement transparency or not.

## **7.5 Conclusion**

The above discussion highlights different factors that may impact the extent to which an ed-tech company might want to make their machine learning powered AI implementations transparent. These factors may determine the different aspects of the Transparency Index Framework that an ed-tech company chooses to implement in their AI development process. The choice of which parts of the Transparency Index Framework are implemented in the AI development process also depends on whether they are being made transparent for ed-tech company's internal use or for external stakeholders. From the viewpoint of the users of an ed-tech tool, there are a number of factors discussed above that may impact the strategic decisions that are

taken by the ed-tech company to determine if the tool is high impact, will have a 'human in the loop' and be used by users who are not tech savvy.

## **Chapter 8: Conclusion and Future Work**

The Transparency Index Framework proposed in this research integrates the popular frameworks used for ethical AI, such as Datasheets, Model Cards and Factsheets into one coherent framework that addresses the whole AI product development timeline and adapts it for the educational technology context. These existing frameworks have been complemented with more requirements like models evaluation report and models validation report to suit AI for educational contexts. The initial version of the Transparency Index Framework that drew on these tools and on a review of the literature was subsequently evaluated with user stakeholder groups and modified to validate the framework's structure and content.

As shown in figure 3a, this research conceptualises transparency for different AI in education stakeholders including educators, ed-tech experts and AI practitioners.

	Main Research Question: What design framework can be applied to ensure optimal level of transparency in machine learning powered Artificial Intelligence (AI) products in educational contexts?			
Research Questions	RQ 1: How should an optimal level of transparency be conceptualised for different stakeholders of machine learning powered AI in Education (AIED)	RQ 2: How can existing frameworks for ethical AI be applied in the context of education for transparent AI development pipelines	RQ 3: How can a design framework assist in minimising and understanding the Awareness Gap in AI powered ed-tech?	RQ 4: How can a design framework be utilised by different stakeholders to make more informed decisions regarding the selection and development of AI powered ed-tech?
Gaps	Gap: Conceptualization of transparency for AI in Education: theoretically and pragmatically for stakeholders like educators, ed-tech experts and AI practitioners	Gap: A theoretical and empirically based framework for transparent machine learning powered AI development pipelines in educational contexts	Gap: Exploration of factors that address the Awareness Gap in educational contexts for AI products	Gap: Evaluation of a design framework for transparency in AI powered ed-tech products through different stakeholders of AI in Education
Goals	Goal: Contextualize transparency for the diverse set of stakeholders of AI in Education including ed-tech users and companies	Goal: Build a framework for transparency in machine learning powered AI products in educational settings, based on existing ethical AI frameworks	Goal: Empirically and theoretically detect the factors from users (learners, educators and ed-tech experts) and suppliers (ed-tech companies) perspective that impact the awareness gap	Goal: Evaluate and improve the framework based on stakeholders' feedback including educators, ed-tech experts and AI practitioners working on ed-tech products
Methodology	Methodology: Thorough literature review of Transparency in AI and Transparency in AI for AIED and interviews with different stakeholders of AI in Education	Methodology: Literature review to identify different ethical AI frameworks and then real-world implementation of these frameworks in educational contexts	Methodology: Practical implementation of ethical AI frameworks in an educational context and interviews with different stakeholders of AI in Education	Methodology: Interviews with educators, ed-tech experts and AI practitioners to incorporate their feedback in Transparency Index framework
Contributions	Contribution: An interpretation of Transparency in accordance with the different tiers of stakeholders of AI in Education	Contribution: Application and alteration of exiting ethical AI frameworks (for different stages of the AI development process) to suit educational contexts	Contribution: A Transparency Index Framework for machine learning powered AI in Education covering the entire AI development pipeline	Contribution: Insights into the requirements of educators, ed-tech experts and AI practitioners for transparent AI powered ed-tech products

Figure 3a: Research question and contributions of this research

The figure shows how domain-agnostic ethical AI frameworks can be applied in education and what measures can be taken in the data processing stage of AI development process to reduce the Awareness Gap of collected digital data and human experiences in the real world.

The Transparency Index is a comprehensive framework that is designed to evaluate, audit and analyse the effectiveness of AI systems in education. It has been designed to benefit different stakeholders of AI in education including educators, teachers, learners, ed-tech experts, executive leaders and AI practitioners developing ed-tech products. Educators can utilize this framework to evaluate the AI-powered ed-tech being used in their schools, AI practitioners can use this as a checklist to document the robustness of their AI development processes and ed-tech experts can use the Transparency Index Framework as an auditing tool before recommending an AI-powered ed-tech product.

Recently, there has been significant research work on developing checklists and frameworks for ethical AI. The research conducted as a part of this thesis takes such work forward by proposing a robust framework for transparency in AI systems applied in education. It shows how AI practitioners and ed-tech companies developing AI-powered products can make sense of the measures they take to ensure ethical AI for different tiers of stakeholders. It also highlights the importance of transparency for companies to develop robust and ethical AI development pipelines and for stakeholders to get a better understanding of how the AI systems that impact them, actually work.

## **8.1 Limitations**

One of the limitations of the Transparency Index Framework is that it was evaluated with a limited number of different stakeholders of AI in education only within the United Kingdom. It is possible for the framework to not work as effectively in other locations like Asia and North America with different regulations on personal data collection, where adoption of AI and ed-tech in schools is not the same as in the UK and the curriculum and culture of schools vary significantly.

Transparency Index Framework is based on assumptions which may not always hold true. For example, one limitation of the framework is in the generalisation of the stakeholders of AI in education in three tiers based on their tech savviness. The boundaries between these tiers can be blurred and at times difficult to differentiate. For example, software engineers are very technical and have an in-depth understanding of how ed-tech products are built but their understanding of machine learning, data engineering or limitations of AI systems can be limited making them suitable for tier 2 stakeholders rather than tier 1.

Number of participants in the interviews is also a limitation and can be increased in future to gather more detailed feedback from different stakeholders. The current version of the Transparency Index Framework for AI in education was developed in two iterations with three different groups of

stakeholders of AI in education including educators, ed-tech experts and AI practitioners. The findings of the framework can be further cemented with more interviews with these stakeholders as well as by including more stakeholders like learners, their parents, regulators and executive leaders of ed-tech companies.

## **8.2 Contribution**

There has been a lot of work conducted to develop checklists and frameworks for ethical AI (Geburu et al, 2018; Mitchell et al, 2019; Bellamy et al, 2019). The research reported here takes such work forward by proposing a robust checklist and a set of guidelines for the transparency of AI systems used in educational settings. It offers a tool for AI practitioners and ed-tech companies developing AI-powered products to make sense of the measures they take to ensure ethical AI for different tiers of stakeholders. It also highlights the importance of Transparency for companies to develop robust AI development pipelines and for stakeholders to get a better understanding of how the AI systems that impact them work.

Research has shown that the framing of educational research evidence in the context of AI affects its perceived credibility (Cukurova et al, 2020), and rightly so. The research presented in this thesis aims to bridge this gap between educators and AI through transparency. It can empower the educators to ask the right questions to get more insights and enhance their understanding of the AI systems. Transparency Index Framework also offers guidelines to ed-tech companies and AI practitioners on how to make their AI systems robust and more accessible for educators and learning sciences community. In essence, the TIF can be considered a step towards bringing the learning sciences and machine learning communities closer by providing the ML community a pathway to develop ed-tech tools whose inner workings and development process are easily understandable by the learning sciences community. The 'Implementation Transparency' section of the Transparency Index Framework can facilitate in incorporating the feedback of learning scientists in developing AI-powered ed-tech tools. The framework also



enables a shared understanding of AI-powered ed-tech tools by different stakeholders of AI in education such as educators, ed-tech experts and AI practitioners.

Different components of TIF mentioned in section 7.2 can be utilized by various stakeholders to enhance their understanding of AI-powered ed-tech tools. The checklist can be used by AI practitioners during AI development to document the different tools used, testing processes followed, decisions taken and assumptions made in the development process. They can also benefit from the second and third components of TIF (stakeholder categorization and requirements for transparency) if they want to make their AI-powered ed-tech tool transparent for a particular category of stakeholders.

Ed-tech experts and educators can also use the first component of TIF (checklist) to audit the AI-powered ed-tech tools before deploying them in their institutions. They can also benefit from the third component of TIF (requirements for transparency) to explore the different ethical considerations taken into account during the AI-powered ed-tech tool's development process.

Implementation of TIF on AI-powered ed-tech products would also enable reproducibility of AIED research, as new researchers willing to replicate or improve the results for ML models used in ed-tech products would have access to the detailed documentation on every aspect of the AI development process.

Another important contribution of TIF is that it brings the diversity and inclusivity limitations of AI systems in educational contexts to the surface, thereby encouraging the ed-tech companies developing these tools to make their AI-powered products more inclusive for a diverse group of stakeholders.

### **8.3 Future Work**

The Transparency Index Framework proposed in this research can be further developed to make it more interpretable for the public. The current form of the

framework was created for AI practitioners, ed-tech experts and educators but it can be further refined to accommodate other stakeholders of AI in education like parents, learners, executives and policy makers.

Based on the checklist and the details proposed in this research, companies can be allocated a Transparency Index Score (TIS) on the number line that highlights a company's score out of hundred. A company score could be based on the measures they have taken (from the Transparency Index Framework) to make the AI development pipeline transparent for various stakeholders. This score would illustrate where the company stands in comparison with a fully transparent AI system. This mapping can make it easier for not just educators but also parents who are tier 3 users (in figure 21) to have a better understanding of the AI-powered ed-tech products their kids are using. The effectiveness of these scores in reflecting the transparency of AI development processes for ed-tech products can be further explored by interviewing different stakeholders like end-users and regulators.

Ed-tech companies can also share the index score with their clients and with the public through their website. This can give their claims about ethical and robust AI more credibility and ease the evaluation process for educators.

In future, the adoption of Transparency Index Framework can also be explored in other industries like healthcare, financial services and law enforcement by engaging with relevant stakeholders. It would be interesting to analyze and evaluate the changes Transparency Index Framework goes through in different sectors with various stakeholders within each sector.

In future, the framework can also be applied and adopted to suit educational institutions in other geographic locations with different regulations, institutional culture and rate of ed-tech adoption. It would be valuable for cross-context validation to research and document the changes the framework goes through in different contexts.

## References

1. A. Dutt, M. A. Ismail and T. Herawan, 2017, "A Systematic Review on Educational Data Mining," in *IEEE Access*, vol. 5, pp. 15991-16005, doi: 10.1109/ACCESS.2017.2654247.
2. Abedjan, Z., 2022. Enabling data-centric AI through data quality management and data literacy. *it-Information Technology*, 64(1-2), pp.67-70.
3. Abràmoff, M.D., Tobey, D. and Char, D.S., 2020. Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *American journal of ophthalmology*, 214, pp.134-142.
4. Abusitta, A., Aïmeur, E. and Wahab, O.A., 2019. Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems. *arXiv preprint arXiv:1905.09972*.
5. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, pp.52138-52160.
6. Agarwal, S., Pandey, G.N. and Tiwari, M.D., 2012. Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 2(2), p.140.
7. Aggarwal, C.C., 2018. Neural networks and deep learning. *Springer*, 10, pp.978-3.
8. Agudo-Peregrina, Á.F., Hernández-García, Á. and Iglesias-Pradas, S., 2012, October. Predicting academic performance with learning analytics in virtual learning environments: A comparative study of three interaction classifications. In *2012 International Symposium on Computers in Education (SIIE)* (pp. 1-6). IEEE.
9. Ahmad, M.A., Eckert, C. and Teredesai, A., 2018, August. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559-560).

10. Aldowah, H., Al-Samarraie, H. and Fauzy, W.M., 2019. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37, pp.13-49.
11. Aljazeera, [www.aljazeera.com](http://www.aljazeera.com). (2021). *Privacy fears as India's gov't schools install facial recognition*. [online] Available at: <https://www.aljazeera.com/news/2021/3/2/privacy-fears-as-indias-govt-schools-install-facial-recognition>
12. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S. and Asari, V.K., 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.
13. Alonso, J.M. and Casalino, G., 2019, June. Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In *International workshop on higher education learning methodologies and technologies online* (pp. 125-138). Springer, Cham.
14. Alwahaby, H., Cukurova, M., Papamitsiou, Z., & Giannakos, M. (2021, August 23). The evidence of impact and ethical considerations of Multimodal Learning Analytics: A Systematic Literature Review. <https://doi.org/10.35542/osf.io/sd23y>
15. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B. and Zimmermann, T., 2019, May. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291-300). IEEE.
16. Amodei, D., Brain, G., Olah, C., Steinhardt, J., Christiano, P., John Schulman Openai and Mané, D. (2016). Concrete Problems in AI Safety. [online] Available at: <https://arxiv.org/pdf/1606.06565v1.pdf>.
17. Ananny, Mike. and Crawford, Kate., (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New media & society*, 20(3):973–989.
18. Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). *Machine Bias*. [online] ProPublica. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

19. Anjaria, K., 2021. A framework for ethical artificial intelligence-from social theories to cybernetics-based implementation. *International Journal of Social and Humanistic Computing*, 4(1), pp.1-28.
20. Apampa, K., G. Wills, and D. Argles. 2010. "An Approach to Presence Verification in Summative e-Assessment Security." In 2010 International Conference on Information Society, 647–651. IEEE.
21. Arnold, M., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K.N., Olteanu, A., Piorkowski, D. and Reimer, D., 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), pp.6-1.
22. Avellan, T., Sharma, S. and Turunen, M., 2020, January. AI for all: defining the what, why, and how of inclusive AI. In *Proceedings of the 23rd International Conference on Academic Mindtrek* (pp. 142-144).
23. Ayodele, T.O., 2010. Types of machine learning algorithms. *New advances in machine learning*, 3, pp.19-48.
24. Baker, R.S. and Hawn, A., 2021. Algorithmic Bias in Education.
25. Balayn, A., Lofi, C. and Houben, G.J., 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, pp.1-30.
26. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. <http://www.fairmlbook.org>
27. Basit, T., 2003. Manual or electronic? The role of coding in qualitative data analysis. *Educational research*, 45(2), pp.143-154.
28. Bellamy, R.K.E., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J. and Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, [online] 63(4/5), pp.4:1-4:15. Available at: <https://ieeexplore.ieee.org/abstract/document/8843908>

29. Belotto, M.J., 2018. Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. *The Qualitative Report*, 23(11), pp.2622-2633.
30. Bennane, A., 2013. Adaptive educational software by applying reinforcement learning. *Informatics in Education-An International Journal*, 12(1), pp.13-27.
31. Bennett, C.L. and Keyes, O., 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125), pp.1-1.
32. Berendt, B., A. Littlejohn, P. Kern, P. Mitros, X. Shacklock, and M. Blakemore. 2017. *Big Data for Monitoring Educational Systems*. Luxembourg: Publications Office of the European Union. <https://publications.europa.eu/en/publication-detail/-/publication/94cb5fc8-473e-11e7-aea8-01aa75ed71a1/>.
33. Bettina Berendt, Allison Littlejohn & Mike Blakemore (2020) AI in education: learner choice and fundamental rights, *Learning, Media and Technology*, 45:3, 312-324, DOI: [10.1080/17439884.2020.1786399](https://doi.org/10.1080/17439884.2020.1786399)
34. Bhatt, U., Andrus, M., Weller, A. and Xiang, A., 2020. Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408*.
35. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. and Eckersley, P., 2020, January. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648-657).
36. Bhavsar, H. and Ganatra, A., 2012. A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(4), pp.2231-2307.
37. Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F. and Verbert, K., 2018, March. Open learner models and learning analytics dashboards: a systematic review. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 41-50).
38. Bogina, V., Hartman, A., Kuflik, T. and Shulner-Tal, A., 2021. Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics. *International Journal of Artificial Intelligence in Education*, pp.1-26.

39. Bostrom, N. and Yudkowsky, E., 2014. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1, pp.316-334.
40. Bracke, P., Datta, A., Jung, C. and Sen, S., 2019. Machine learning explainability in finance: an application to default risk analysis.
41. Brendel, A.B., Mirbabaie, M., Lembcke, T.B. and Hofeditz, L., 2021. Ethical management of artificial intelligence. *Sustainability*, 13(4), p.1974.
42. Bresfelean, V.P., Bresfelean, M., Ghisoiu, N. and Comes, C.A., 2008, June. Determining students' academic failure profile founded on data mining methods. In *ITI 2008-30th International Conference on Information Technology Interfaces* (pp. 317-322). IEEE.
43. Bridgeman, B., Trapani, C., & Attali, Y. (2009, April 13-17). Considering fairness and validity in evaluating automated scoring [Paper presentation]. Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA, United States.
44. Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, 25(1), 27–40.  
<https://doi.org/10.1080/08957347.2012.635502>
45. Brill J (2015) Scalable approaches to transparency and accountability in decisionmaking algorithms: remarks at the NYU conference on algorithms and accountability. Federal Trade Commission, 28 February. Available at:  
[https://www.ftc.gov/system/files/documents/public\\_statements/629681/150228nyualgorithms.pdf](https://www.ftc.gov/system/files/documents/public_statements/629681/150228nyualgorithms.pdf)
46. Brown, S. P., Cron, W. L., & Slocum Jr, J. W. (1998). Effects of trait competitiveness and perceived intraorganizational competition on salesperson goal setting and performance. *Journal of Marketing*, 62(4), 88-98.
47. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901.

48. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R. and Maharaj, T., 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.
49. Buckley, B. and Hunter, M., 2011. Say cheese! Privacy and facial recognition. *Computer Law & Security Review*, 27(6), pp.637-640.
50. Buhrmester, V., Münch, D. and Arens, M., 2019. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint arXiv:1911.12116*.
51. Bull, S. and Kay, J., 2007. Student models that invite the learner in: The SMILI:( ) Open learner modelling framework. *International Journal of Artificial Intelligence in Education*, 17(2), pp.89-120.
52. Bull, S., 2020. There are open learner models about!. *IEEE Transactions on Learning Technologies*, 13(2), pp.425-448.
53. Burchinal, M., Zaslow, M. and Tarullo, L. eds., 2016. *Quality thresholds, features, and dosage in early care and education: secondary data analyses of child outcomes* (Vol. 81). San Fransisco, CA: Wiley.
54. Burkart, N. and Huber, M.F., 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, pp.245-317.
55. Burrell, J., 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), p.2053951715622512.
56. Caliskan A, Bryson JJ, Narayanan A. 2017 Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186.
57. Callaway, E. (2020). “It will change everything”: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature*. [online] Available at: <https://www.nature.com/articles/d41586-020-03348-4>.
58. Cam, E. and Ozdag, M.E., 2021. Discovery of Course Success Using Unsupervised Machine Learning Algorithms. *Malaysian Online Journal of Educational Technology*, 9(1), pp.26-47.



59. Carabantes M (2019) Black-box artificial intelligence: an epistemological and critical analysis. *AI Soc.* <https://doi.org/10.1007/s00146-019-00888-w>
60. Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R.D.P., Basto, J.P. and Alcalá, S.G., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, p.106024.
61. Castelvechi, D. (2016). Can we open the black box of AI? *Nature*, [online] 538(7623), pp.20–23. Available at: <https://www.nature.com/articles/doi:10.1038/538020a>
62. Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. ArXiv E-Prints, arXiv:2010.04053. <https://arxiv.org/abs/2010.04053>
63. Çetinkaya, Y.M., Toroslu, İ.H. and Davulcu, H., 2020. Developing a Twitter bot that can join a discussion using state-of-the-art architectures. *Social network analysis and mining*, 10(1), pp.1-21.
64. Chambers S (2004) Behind closed doors: publicity, secrecy, and the quality of deliberation. *J Polit Philos* 12(4):389–410
65. Chambers S (2005) Measuring publicity's effect: reconciling empirical research and normative theory. *Acta Polit* 40(2):255–266
66. Chan, A., Okolo, C.T., Ternier, Z. and Wang, A., 2021. The Limits of Global Inclusion in AI Development. arXiv preprint arXiv:2102.01265.
67. Chander, A., Srinivasan, R., Chelian, S., Wang, J. and Uchino, K., 2018, January. Working with beliefs: AI transparency in the enterprise. In *IUI Workshops*.
68. Chang, T.C. and Wang, H., 2016. A multi criteria group decision-making model for teacher evaluation in higher education based on cloud model and decision tree. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(5), pp.1243-1262.
69. Chaudhry, M.A., Cukurova, M. and Luckin, R., 2022. A Transparency Index Framework for AI in Education.
70. Chen, C.M., Chen, Y.Y. and Liu, C.Y., 2007. Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6), pp.1349-1359.

71. Chen, F., 2022. Human-AI Cooperation in Education: Human in Loop and Teaching as leadership. *Journal of Educational Technology and Innovation* 本刊已被维普网全文收录, 2(01).
72. Chen, K., Hellerstein, J.M. and Parikh, T.S., 2011, January. Data in the First Mile. In *CIDR* (pp. 203-206).
73. Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163.
74. Christie, S. T., Jarratt, D. C., Olson, L. A., & Tajjala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, 726–731.
75. Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.
76. Corbett-Davies, S.; Goel, S.; Morgenstern, J.; and Cummings, R. 2018. Defining and designing fair algorithms. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 705–705. ACM.
77. Cox, A.M., 2021. Exploring the impact of Artificial Intelligence and robots on higher education through literature-based design fictions. *International Journal of Educational Technology in Higher Education*, 18(1), pp.1-19.
78. Cox, D.R., 2006. Frequentist and Bayesian statistics: A critique (Keynote address). In *Statistical problems in particle physics, astrophysics and cosmology* (pp. 3-6).
79. Cramer, H., Holstein, K., Vaughan, J. W., Daumé, H., Dudik, M., Wallach, H., Reddy, S., & Jean, G.-G. [The Conference on Fairness, Accountability, and Transparency (FAT\*)]. (2019, February 23). *FAT\* 2019 Translation Tutorial: Challenges of incorporating algorithmic fairness* [Video]. YouTube. <https://youtu.be/UicKZv93SOY>
80. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A., 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1), pp.53-65.

81. Cui, Z. and Gong, G., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage*, 178, pp.622-637.
82. Cukurova, M., Luckin, R. and Kent, C., 2020. Impact of an artificial intelligence research frame on the perceived credibility of educational research evidence. *International Journal of Artificial Intelligence in Education*, 30(2), pp.205-235.
83. Cukurova, M., Luckin, R., Clark-Wilson, A., Moore, G., Olatunji, T. and McDonald, M., 2018, June. EDUCATE: Creating the Golden Triangle for Research-Informed Industrial Collaborations within Educational Technology. In *CEUR Workshop Proceedings (Vol. 2128)*. CEUR Workshop Proceedings.
84. Cukurova, M., Luckin, R., Millán, E. and Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, 116, pp.93–109.
85. Cukurova, M., Zhou, Q., Spikol, D. and Landolfi, L., 2020, March. Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough?. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 270-275).
86. Cunningham, P., Cord, M. and Delany, S.J., 2008. Supervised learning. In *Machine learning techniques for multimedia* (pp. 21-49). Springer, Berlin, Heidelberg.
87. Custer, S., E. M. King, and T. M. Atnic. 2018. *Toward Data-Driven Education Systems: Insights into Using Information to Measure Results and Manage Change*. Brookings Institution. Accessed 25 May 2020. <https://www.brookings.edu/research/toward-data-driven-education-systems-insights-into-using-information-to-measure-results-and-manage-change/0>.
88. Cyr, J., 2016. The pitfalls and promise of focus groups as a data collection method. *Sociological methods & research*, 45(2), pp.231-259.
89. Daas, P. and Arends-Tóth, J., 2012. Secondary data collection. *Statistics Netherlands. The Hague*.

90. Dal Pozzolo, A., Caelen, O. and Bontempi, G., 2010. Comparison of balancing techniques for unbalanced datasets. *Mach. Learn. Gr. Univ. Libr. Bruxelles Belgium*, 16(1), pp.732-735.
91. Dameski, A., 2018, August. A comprehensive ethical framework for AI entities: Foundations. In *International Conference on Artificial General Intelligence* (pp. 42-51). Springer, Cham.
92. Danaher, J., 2018. Toward an ethics of AI assistants: An initial framework. *Philosophy & Technology*, 31(4), pp.629-653.
93. Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. [online] U.S. Available at: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
94. Davidson, T., 2019. Black-box models and sociological explanations: Predicting high school grade point average using neural networks. *Socius*, 5, p.2378023118817702.
95. De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92(5), 938–956. <https://doi.org/10.1037/0022-3514.92.5.938>
96. de Fine Licht, K. and de Fine Licht, J., 2020. Artificial intelligence, transparency, and public decision-making. *AI & SOCIETY*, pp.1-10.
97. De Laat PB (2018) Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability? *Philos Technol* 31(4):525–541
98. Demajo, L.M., Vella, V. and Dingli, A., 2020. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*.
99. Demartini, G., Roitero, K. and Mizzaro, S., 2021. Managing Bias in Human-Annotated Data: Moving Beyond Bias Removal. *arXiv preprint arXiv:2110.13504*.
100. Dewan, M., A. Akber, M. Murshed, and F. Lin. 2019. "Engagement Detection in Online Learning: A Review." *Smart Learning Environments* 6 (1): 1. doi: 10.1186/s40561-018-0080-z

101. Diakopoulos N (2016) Accountability in algorithmic decision making. *Communications of the ACM* 59(2): 56–62.
102. Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp.78-87.
103. Dorça, F.A., Lima, L.V., Fernandes, M.A. and Lopes, C.R., 2013. Comparing strategies for modelling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis. *Expert Systems with Applications*, 40(6), pp.2092-2101.
104. Dorodchi, M., Al-Hossami, E., Benedict, A. and Demeter, E., 2019, December. Using Synthetic Data Generators to Promote Open Science in Higher Education Learning Analytics. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 4672-4675). IEEE.
105. Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1702.08608>.
106. Dubber, M.D., Pasquale, F. and Das, S. eds., 2020. *The Oxford Handbook of Ethics of AI*. Oxford University Press, USA.
107. Durmus, M., 2022. COGNITIVE BIASES-A Brief Overview of Over 160 Cognitive Biases:+ Bonus Chapter: Algorithmic Bias.
108. Eitel-Porter, R., 2021. Beyond the promise: implementing ethical AI. *AI and Ethics*, 1(1), pp.73-80.
109. El Guabassi, I., Bousalem, Z., Marah, R. and Qazdar, A., 2021. A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms.
110. Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G. and Rangwala, H., 2016. Predicting student performance using personalized analytics. *Computer*, 49(4), pp.61-69.
111. Elster J (1998) Deliberation and constitution making. In: Elster J (ed) *Deliberative Democracy*. Cambridge University Press, Cambridge
112. Eppler MJ, Mengis J (2004) The concept of overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. *Inf Soc* 20(5):325–344

113. Fan, H., Du, D., Wen, L., Zhu, P., Hu, Q., Ling, H., Shah, M., Pan, J., Schumann, A., Dong, B. and Stadler, D., 2020, August. VisDrone-MOT2020: The Vision Meets Drone Multiple Object Tracking Challenge Results. In *European Conference on Computer Vision* (pp. 713-727). Springer, Cham.
114. Fedus, W., Zoph, B. and Shazeer, N., 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.
115. Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A., 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, pp.1-29.
116. Fernández, R.R., De Diego, I.M., Aceña, V., Fernández-Isabel, A. and Moguerza, J.M., 2020. Random forest explainability using counterfactual sets. *Information Fusion*, 63, pp.196-207.
117. Fiok, K., Farahani, F.V., Karwowski, W. and Ahram, T., 2021. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, p.15485129211028651.
118. Floridi, L., 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6), pp.261-262.
119. Fosch-Villaronga, E. and Poulsen, A., 2022. Diversity and Inclusion in Artificial Intelligence. *Law and Artificial Intelligence*, pp.109-134.
120. Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42.  
<https://doi.org/10.1257/089533005775196732>
121. Fuchs, D.J., 2018. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1), p.1.
122. Fwa, H.L. and Marshall, L., 2018. Modeling engagement of programming students using unsupervised machine learning technique. *GSTF Journal on Computing*.
123. Gade, K., Geyik, S.C., Kenthapadi, K., Mithal, V. and Taly, A., 2019, July. Explainable AI in industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3203-3204).

124. Gallagher, M., 2009. Data collection and analysis. *Researching with children and young people: Research design, methods and analysis*, pp.65-127.
125. Garbin, C. and Marques, O., 2022. Assessing Methods and Tools to Improve Reporting, Increase Transparency, and Reduce Failures in Machine Learning Applications in Health Care. *Radiology: Artificial Intelligence*, 4(2).
126. Garfinkel, S., Matthews, J., Shapiro, S.S. and Smith, J.M., 2017. Toward algorithmic transparency and accountability.
127. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K., 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
128. Gerke, S., Minssen, T. and Cohen, G., 2020. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295-336). Academic Press.
129. Gil, P.D., da Cruz Martins, S., Moro, S. and Costa, J.M., 2021. A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, 26(2), pp.2165-2190.
130. Gilpin, LH, Bau, D, Yuan, BZ, et al. 2018, Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, pp.80–89.
131. Gitelman, L., Jackson, V., Rosenberg, D., Williams, T.D., Brine, K.R., Poovey, M., Stanley, M., Garvey, E.G., Krajewski, M., Raley, R. and Ribes, D., 2013. Data flakes: An afterword to “raw data” is an oxymoron.
132. Goddard, K., Roudsari, A. and Wyatt, J.C., 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), pp.121-127.
133. Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.
134. Greene, D., Hoffmann, A. L. & Stark, L. Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and

- machine learning. In *Proc. 52nd Hawaii International Conference on System Sciences* 2122–2131 (HICSS, 2019).
135. Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., & Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).
  136. Gross, J. J., & John, O. P. (2003). Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being. *Journal of Personality and Social Psychology*, 85(2), 348–362. <https://doi.org/10.1037/0022-3514.85.2.348>
  137. Gruntiz, M., WeAreBrain Blog. (2021). Rule-based AI vs machine learning: what's the difference? [online] Available at: <https://wearebrain.com/blog/ai-data-science/rule-based-ai-vs-machine-learning-whats-the-difference/>.
  138. Gunawardana, A. and Shani, G., 2009. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(12).
  139. Hagendorff, T., 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), pp.99-120.
  140. Halde, R.R., 2016, September. Application of Machine Learning algorithms for betterment in education system. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)* (pp. 1110-1114). IEEE.
  141. Hao, K. (2019). *We analyzed 16,625 papers to figure out where AI is headed next*. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/>.
  142. Hao, K. and Stray, J., 2019. Can you make AI fairer than a judge? Play our courtroom algorithm game. *MIT Technology Review*.
  143. Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315– 3323.



144. Hastie, T., Tibshirani, R. and Friedman, J., 2009. Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41). Springer, New York, NY.
145. Hayes-Roth, F., 1985. Rule-based systems. *Communications of the ACM*, 28(9), pp.921-932.
146. Hayman, B., Wilkes, L., Jackson, D. and Halcomb, E., 2012. Story-sharing as a method of data collection in qualitative research. *Journal of Clinical Nursing*, 21(1-2), pp.285-287.
147. He, H. and Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), pp.1263-1284.
148. Heald D (2006) Varieties of transparency. *Proceedings of the British Academy* 135: 25–43.
149. Herold, B. (2017). The Case(s) Against Personalized Learning. *Education Week*, 37(12), 4-5. Retrieved from <https://www.edweek.org/ew/articles/2017/11/08/the-cases-againstpersonalized-learning.html>
150. Hollanek, T. (2020). AI transparency: a matter of reconciling design with critique. *AI & SOCIETY*.
151. Holzinger, A., 2018, August. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)* (pp. 55-66). IEEE.
152. Holzinger, A., Kieseberg, P., Weippl, E. and Tjoa, A.M., 2018, August. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 1-8). Springer, Cham.
153. Hosny, K.M., Kassem, M.A. and Fouad, M.M., 2020. Classification of skin lesions into seven classes using transfer learning with AlexNet. *Journal of digital imaging*, 33(5), pp.1325-1334.
154. Hox, J.J. and Boeije, H.R., 2005. Data collection, primary versus secondary. <https://doi.org/10.3389/fdata.2019.00013>
155. Hu, B., 2017, May. Teaching quality evaluation research based on neural network for university physical education. In *2017 International*

- Conference on Smart Grid and Electrical Automation (ICSGEA)* (pp. 290-293). IEEE.
156. Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 431–437.
  157. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.
  158. Ibert, J., Baumard, P., Donada, C. and Xuereb, J.M., 2001. Data collection and managing the data source. *RA Thiétart (éds.), Doing management research, a comprehensive guide, Thousand Oaks, Sage Publication*, pp.289-329.
  159. Iglesias, A., Martínez, P., Aler, R. and Fernández, F., 2009. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, 31(1), pp.89-106.
  160. Iglesias, A., Martínez, P., Aler, R. and Fernández, F., 2009. Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowledge-Based Systems*, 22(4), pp.266-270.
  161. Irving, G. and Askeel, A., 2019. AI safety needs social scientists. *Distill*, 4(2), p.e14.
  162. Irving, G., Christiano, P. and Amodei, D., 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.
  163. Ivančević, V., Čeliković, M. and Luković, I., 2012, October. The individual stability of student spatial deployment and its implications. In *2012 International Symposium on Computers in Education (SIIE)* (pp. 1-4). IEEE.
  164. Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R. and Sycara, K. (2018). Transparency and Explanation in Deep Reinforcement Learning Neural Networks. [online] arXiv.org. Available at: <https://arxiv.org/abs/1809.06061>.
  165. Jack, K. 2018. How AI Can Spot Exam Cheats and Raise Standards. *Financial Times* (London). 20 Aug 2018. Accessed 25 May 2020. <https://www.ft.com/content/540e77fa-9fe2-11e8-85da-eeb7a9ce36e4>.

166. Jackson, S. and Panteli, N., 2021, September. A Multi-level Analysis of Mistrust/Trust Formation in Algorithmic Grading. In *Conference on e-Business, e-Services and e-Society* (pp. 737-743). Springer, Cham.
167. Jacobson, M. et al. (2019) Education as a complex system: conceptual and methodological implications *Educational Researcher*, Vol. 48, No. 2
168. Jain, N., 2021. Survey Versus Interviews: Comparing Data Collection Tools for Exploratory Research. *The Qualitative Report*, 26(2), pp.541-554.
169. Jameel, S.M., Hashmani, M.A., Alhussain, H., Rehman, M. and Budiman, A., 2020. A critical review on adverse effects of concept drift over machine learning classification models. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(1), p.2020.
170. Januszewski, A. and Molenda, M. eds., 2013. *Educational technology: A definition with commentary*. Routledge.
171. Jars-Quant, <https://apastyle.apa.org>. (2018). Quantitative research design (JARS–Quant). [online] Available at: <https://apastyle.apa.org/jars/quantitative>.
172. Jo, E.S. and Gebru, T., 2020, January. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 306-316).
173. Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp.389-399.
174. Johnston, M.P., 2017. Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries*, 3(3), pp.619-626.
175. Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp.255-260.
176. Jotterand, F. and Bosco, C., 2020. Keeping the “human in the loop” in the age of artificial intelligence. *Science and Engineering Ethics*, 26(5), pp.2455-2460.
177. Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4, pp.237-285.
178. Kai, S., Andres, J. M. L. ., Paquette, L., Baker, R. S. ., Molnar, K., Watkins, H., & Moore, M. (2017). Predicting Student Retention from

- Behavior in an Online Orientation Course. Proceedings of the 10th International Conference on Educational Data Mining, 250–255.
179. Karam, R., Pane, J. F., Griffin, B. A., Robyn, A., Phillips, A., & Daugherty, L. (2017). Examining the implementation of technology-based blended algebra I curriculum at scale. *Educational Technology Research & Development*, 65, 399-425. doi:<https://doi.org/10.1007/s11423-016-9498-6>
  180. Kardan, A.A., Sadeghi, H., Ghidary, S.S. and Sani, M.R.F., 2013. Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, 65, pp.1-11.
  181. Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).
  182. Katal, A., Wazid, M. and Goudar, R.H., 2013, August. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)* (pp. 404-409). IEEE.
  183. Kay, J., 2012. AI and education: grand challenges. *IEEE Intelligent Systems*, 27(5), pp.66-69.
  184. Kazim, E. and Koshiyama, A.S. (2021). A high-level overview of AI ethics. *Patterns*, 2(9), p.100314.
  185. Kazim, E., 2017. *Kant on conscience: A unified approach to moral self-consciousness*. Brill.
  186. Kent, C. and Cukurova, M., 2020. Investigating collaboration as a process with theory-driven learning analytics. *Journal of Learning Analytics*, 7(1), pp.59-71.
  187. Khanum, M., Mahboob, T., Imtiaz, W., Ghafoor, H.A. and Sehar, R., 2015. A survey on unsupervised machine learning algorithms for automation, classification and maintenance. *International Journal of Computer Applications*, 119(13).
  188. Kilkeny, M.F. and Robinson, K.M., 2018. Data quality: “Garbage in—garbage out”.
  189. Kim, B., Park, J. and Suh, J., 2020. Transparency and accountability in AI decision support: Explaining and visualizing

- convolutional neural networks for text information. *Decision Support Systems*, 134, p.113302.
190. Kim, B., Suh, H., Heo, J. and Choi, Y., 2020. AI-Driven Interface Design for Intelligent Tutoring System Improves Student Engagement. *arXiv preprint arXiv:2009.08976*.
  191. Kippin, S. and Cairney, P., 2021. The COVID-19 exams fiasco across the UK: four nations and two windows of opportunity. *British Politics*, pp.1-23.
  192. Kizilcec, R. F., & Lee, H. (2020). Algorithmic Fairness in Education. ArXiv E-Prints, arXiv:2007.05443. <https://arxiv.org/abs/2007.05443>
  193. Klugman, C.M., 2021. Black boxes and bias in AI challenge autonomy. *The American Journal of Bioethics*, 21(7), pp.33-35.
  194. Knox, J., Wang, Y. and Gallagher, M., 2019. Introduction: AI, inclusion, and 'everyone learning everything'. In *Artificial Intelligence and Inclusive Education* (pp. 1-13). Springer, Singapore.
  195. Kobayashi, Y., Ishibashi, M. and Kobayashi, H., 2019. How will "democratization of artificial intelligence" change the future of radiologists?. *Japanese journal of radiology*, 37(1), pp.9-14.
  196. Kolbjørnsrud, V., Amico, R. and Thomas, R.J., 2016. How artificial intelligence will redefine management. *Harvard Business Review*, 2(1), pp.3-10.
  197. Kolkman, D., 2020. F\*\* k the algorithm?: what the world can learn from the UK's A-level grading fiasco. *Impact of Social Sciences Blog*.
  198. Kollias, D., Cheng, S., Ververas, E., Kotsia, I. and Zafeiriou, S., 2020. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5), pp.1455-1484.
  199. Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S. and Lomas, E. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. [online] [papers.ssrn.com](https://papers.ssrn.com). Available at: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3778998](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998).

200. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), pp.3-24.
201. Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, [online] 25, pp.1097–1105. Available at:  
<https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>.
202. Kucak, D., Juričić, V. and Đambić, G., 2018. MACHINE LEARNING IN EDUCATION- A SURVEY OF CURRENT RESEARCH TRENDS. *Annals of DAAAM & Proceedings*, 29.
203. Kyriakou, K., Barlas, P., Kleanthous, S. and Otterbacher, J., 2019, July. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAI Conference on Web and Social Media* (Vol. 13, pp. 313-322).
204. Larsson, S. and Heintz, F., 2020. Transparency in artificial intelligence. *Internet Policy Review*, 9(2).
205. Leahy, S.M., Holland, C. and Ward, F., 2019. The digital frontier: Envisioning future technologies impact on the classroom. *Futures*, 113, p.102422.
206. Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. In B. Goertzel, & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006* (pp. 17–24). IOS Press.
207. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2017) Fair, transparent, and accountable algorithmic decision-making processes. *Philos Technol* 2017:1–17
208. Li, S., Chen, G., Xing, W., Zheng, J. and Xie, C., 2020. Longitudinal clustering of students' self-regulated learning behaviors in engineering design. *Computers & Education*, 153, p.103899.
209. Liang, Y., Li, S., Yan, C., Li, M. and Jiang, C., 2021. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419, pp.168-182.

210. Liaw, H., M. Chiu, and C. Chou. 2014. "Using Facial Recognition Technology in the Exploration of Student Responses to Conceptual Conflict Phenomenon." *Chemistry Education Research and Practice* 15 (4): 824–834. doi: 10.1039/C4RP00103F
211. Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2021. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), p.18.
212. Lipton, ZC . The mythos of model interpretability. *ACM Queue* 2018; 61: 36–43.
213. Liu, M.C., Lai, C.H., Su, Y.N., Huang, S.H., Chien, Y.C., Huang, Y.M. and Hwang, J.P., 2015. Learning with great care: The adoption of the multi-sensor technology in education. In *Sensing technology: Current status and future trends III* (pp. 223-242). Springer, Cham.
214. Lohr, S., 2018. Facial recognition is accurate, if you're a white guy. *New York Times*, 9(8), p.283.
215. Long, P., and G. Siemens. 2011. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE Review* 46 (5): 30–40. (September/October 2011). Accessed 25 May 2020. <http://er.educause.edu/articles/2012/7/~link.aspx?id=82AE6F528BDC4EBFBC17FCD75B3E3E5C&z=z>. [Google Scholar]; Wolff et al. 2013; Nistor, Derntl, and Klamma 2015; Papamitsiou and Economides 2014)
216. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J. and Zhang, G., 2018. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), pp.2346-2363.
217. Luan, H. and Tsai, C.C., 2021. A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1), pp.250-266.
218. Luckin, R., Holmes, W., Griffiths, M. and Forcier, L.B., 2016. Intelligence unleashed: An argument for AI in education.
219. Luckin, R., Underwood, J., du Boulay, B., Holmberg, J., Kerawalla, L., O'Connor, J., Smith, H. and Tunley, H. (2006). Designing Educational Systems Fit for Use: A Case Study in the Application of Human Centred Design for AIED. *International Journal of Artificial Intelligence in Education*, [online] 16(4), pp.353–380. Available at:

- <https://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai16-4-03> [Accessed 5 Nov. 2021].
220. Lum, K., 2017. Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour*, 1(7), pp.1-1.
221. Madaio, M.A., Stark, L., Wortman Vaughan, J. and Wallach, H., 2020, April. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).
222. Madhulatha, T.S., 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.
223. Mahapatra, S.S. and Khan, M.S., 2007. A neural network approach for assessing quality in technical education: an empirical study. *International Journal of Productivity and Quality Management*, 2(3), pp.287-306.
224. Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90, pp.46–60.
225. Mangal, S.K. and Mangal, U., 2019. *Essentials of educational technology*. PHI Learning Pvt. Ltd..
226. Mangaroska, K. and Giannakos, M., 2018. Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, s2(4), pp.516-534.
227. Marcus, G., 2020. The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
228. Marda, V., 2018. Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), p.20180087.
229. Mark, M. M., & Shotland, R. (1987). *Multiple methods in program evaluation*. San Francisco, CA: Jossey-Bass. <https://doi.org/10.1002/ev.1461>
230. Martens, B., 2018. The importance of data access regimes for artificial intelligence and machine learning.
231. Martinez, D., Malyska, N., Streilein, B., Reynolds, D., Campbell, W., Richardson, F., Dagli, C., Sahin, C., Gadepally, V., Tran, A., Greenfield, K., Trepagnier, P., Hall, R., Zipkin, J., Roeser, C., Mohindra, S.,



- Hennighausen, K. and Thornton, J. (2019). *Artificial Intelligence: Short History, Present Developments, and Future Outlook Final Report*. [online]. Available at: [https://www.ll.mit.edu/sites/default/files/publication/doc/2021-03/Artificial%20Intelligence%20Short%20History%2C%20Present%20Developments%2C%20and%20Future%20Outlook%20-%20Final%20Report%20-%202021-03-16\\_0.pdf](https://www.ll.mit.edu/sites/default/files/publication/doc/2021-03/Artificial%20Intelligence%20Short%20History%2C%20Present%20Developments%2C%20and%20Future%20Outlook%20-%20Final%20Report%20-%202021-03-16_0.pdf).
232. Maseleno, A., Sabani, N., Huda, M., Ahmad, R.B., Jasmi, K.A. and Basiron, B., 2018. Demystifying learning analytics in personalised learning. *International Journal of Engineering and Technology (UAE)*.
233. Matetic, M., 2019, May. Mining Learning Management System Data Using Interpretable Neural Networks. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1282-1287). IEEE.
234. Mayo, D.G. and Cox, D.R., 2006. Frequentist statistics as a theory of inductive inference. In *Optimality* (pp. 77-97). Institute of Mathematical Statistics.
235. McDermid, J.A., Jia, Y., Porter, Z. and Habli, I., 2021. Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A*, 379(2207), p.20200363.
236. Meeki, N., Amine, A., Boudia, M.A. and Hamou, R.M., 2020. Deep learning for non verbal sentiment analysis: Facial emotional expressions. *GeCoDe Laboratory, Department of Computer Science, Tahar Moulay University of Saida*.
237. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), pp.1-35.
238. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), pp.1-35.
239. Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

240. Meta. (2021). *An Update On Our Use of Face Recognition*. [online] Available at: <https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/>
241. Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, 9, 70–89.
242. Minnaar, L. and Heystek, J., 2013. Online surveys as data collection instruments in education research: A feasible option?. *South African Journal of Higher Education*, 27(1), pp.162-183.
243. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019, January. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
244. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8. <https://doi.org/10.1146/annurev-statistics-042720-125902>
245. Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), pp.501-507.
246. Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the conference on fairness, accountability, and transparency - FAT\* '19*, 279–288. <https://doi.org/10.1145/3287560.3287574>
247. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M., 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
248. Mohamed, A.E., 2017. Comparative study of four supervised machine learning techniques for classification. *International Journal of Applied*, 7(2).
249. Molenaar, I., 2022. The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence*, 3, p.100070.
250. Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L. and Ghazi Saeedi, M., 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and

- evaluation methods. *Interactive Learning Environments*, 29(1), pp.142-163.
251. Muchlinski, D., Siroky, D., He, J. and Kocher, M., 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1), pp.87-103.
252. Mueller, ST, Hoffman, RR, Clancey, W, et al. February 2019; Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv e-prints. arXiv:1902.01876.
253. Muers, S., 2020. Lessons from the A-levels fiasco: putting culture and values at the heart of policymaking. *British Politics and Policy at LSE*.
254. Mun, S.K., Wong, K.H., Lo, S.C.B., Li, Y. and Bayarsaikhan, S., 2021. Artificial Intelligence for the Future Radiology Diagnostic Service. *Frontiers in Molecular Biosciences*, 7, p.512.
255. Naurin D (2007) *Deliberation behind closed doors: transparency and lobbying in the European Union*. ECPR Press, Colchester
256. Neff G, Nagy P (2016) Automation, algorithms, and politics| talking to bots: symbiotic agency and the case of tay. *Int J Commun* 10:17
257. Neto, M.P. and Paulovich, F.V., 2020. Explainable Matrix-Visualization for Global and Local Interpretability of Random Forest Classification Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), pp.1427-1437.
258. Nistor, N., M. Derntl, and R. Klamma. 2015. "Learning Analytics: Trends and Issues of the Empirical Research of the Years 2011-2014." In *Design for Teaching and Learning in a Networked World. EC-TEL 2015*, edited. by G. Conole, T. Klobucar, C. Rensing, J. Konert, and É Lavoué, 453–459. Berlin etc.: Springer. LNCS 9307.
259. Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E. and Kompatsiaris, I., 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), p.e1356.
260. Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of

- populations. *Science*, [online] 366(6464), pp.447–453. Available at: <https://science.sciencemag.org/content/366/6464/447.full>.
261. Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487–501. <https://www.learntechlib.org/p/148344>
262. Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2, 13.
263. Ortega, P.A., Miani, V. and Research, Deepmind Safety. (2018). *Building safe artificial intelligence: specification, robustness, and assurance*. [online] Medium. Available at: <https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1>
264. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), pp.128-138.
265. Páez, A . Mach 2019 The pragmatic turn in explainable artificial intelligence (XAI). *Mind*; 29: 441–59.
266. Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, 36(2), 127- 144.  
doi:<http://dx.doi.org/10.3102/0162373713507480>
267. Papamitsiou, Z., and A. A. Economides. 2014. “Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence.” *Journal of Educational Technology & Society* 17 (4): 49–64.
268. Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who’s Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, 12(3), 1–30.  
<https://doi.org/10.5281/zenodo.4143612>
269. Parajuli, B.K., 2004. Questionnaire: A Tool of Primary Data Collection. *Himalayan Journal of Sociology and Anthropology*, 1, pp.51-63.

270. Parsheera, S., 2019. Adoption and Regulation of Facial Recognition Technologies in India: Why and Why Not?.
271. Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
272. Patel, K., Mehta, D., Mistry, C., Gupta, R., Tanwar, S., Kumar, N. and Alazab, M., 2020. Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. *IEEE Access*, 8, pp.90495-90519.
273. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). Data and its
274. Peña, A., Serna, I., Morales, A. and Fierrez, J., 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 28-29).
275. Perez, S. (2016). *Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]*. [online] TechCrunch. Available at: <https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/>.
276. Petkovic, D., Alavi, A., Cai, D. and Wong, M., 2021. Random Forest Model and Sample Explainer for Non-experts in Machine Learning—Two Case Studies. *Lect. Notes Comput. Sci.*, pp.62-75.
277. Petkovic, D., Altman, R., Wong, M. and Vigil, A., 2018. Improving the explainability of Random Forest classifier—user centered approach. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* (pp. 204-215).
278. Petkovic, D., Altman, R., Wong, M. and Vigil, A., 2018. Improving the explainability of Random Forest classifier—user centered approach. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* (pp. 204-215).
279. Phillips, P. J.; Jiang, F.; Narvekar, A.; Ayyad, J.; and O'Toole, A. J. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)* 8(2):14.
280. Pienta, A.M., O'Rourke, J.M. and Franks, M.M., 2011. Getting started: Working with secondary data. *Secondary data analysis: An introduction for psychologists*, pp.13-25.

281. Plale, B., 2019, September. Transparency by Design in eScience Research. In *2019 15th International Conference on eScience (eScience)* (pp. 428-431). IEEE.
282. Polson, M.C. and Richardson, J.J., 2013. *Foundations of intelligent tutoring systems*. Psychology Press.
283. Ponce, O.A. and Pagán-Maldonado, N., 2015. Mixed methods research in education: Capturing the complexity of the profession. *International journal of educational excellence*, 1(1), pp.111-135.
284. Prain, V., Cox, P., Deed, C., Dorman, J., Edwards, D., Farrelly, C., Keeffe, M., Lovejoy, V., Mow, L., Sellings, P. and Waldrip, B., 2013. Personalised learning: Lessons to be learnt. *British Educational Research Journal*, 39(4), pp.654-676.
285. Pringle, R., K. Michael, and M. G. Michael. 2016. "Unintended Consequences of Living with AI." *IEEE Technology and Society Magazine* 35 (4): 17–21.
286. Prinsloo, P., and S. Slade. 2016. "Student Vulnerability and Agency in Networked, Digital Learning." *European Journal of Open, Distance and E-Learning* 19 (2): 14–34.
287. Prinsloo, P., and S. Slade. 2017. An Elephant in the Learning Analytics Room: The Obligation to Act. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 46–55). ACM.
288. Prunkl, C. and Whittlestone, J., 2020, February. Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 138-143).
289. Purba, W., Tamba, S. and Saragih, J., 2018, April. The effect of mining data k-means clustering toward students profile model drop out potential. In *Journal of Physics: Conference Series* (Vol. 1007, No. 1, p. 012049). IOP Publishing.
290. Putnam, V. and Conati, C., 2019, March. Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). In *IUI Workshops* (Vol. 19, pp. 1-7).
291. Qiu, L. and Riesbeck, C.K., 2008. Human-in-the-loop: a feedback-driven model for authoring knowledge-based interactive learning

- environments. *Journal of Educational Computing Research*, 38(4), pp.469-509.
292. Quadri, M.M. and Kalyankar, N.V., 2010. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*.
293. Rai, A., 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), pp.137-141.
294. Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25–39. <https://doi.org/https://doi.org/10.1016/j.asw.2012.10.004>
295. Rao, T.R., Mitra, P., Bhatt, R. and Goswami, A., 2019. The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, 60(3), pp.1165-1245.
296. Ravi, S. and Larochelle, H., 2016. Optimization as a model for few-shot learning.
297. Ray, S., 2019, February. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.
298. Re, R.M. and Solow-Niederman, A., 2019. Developing artificially intelligent justice. *Stan. Tech. L. Rev.*, 22, p.242.
299. Reddi, V.J., Damos, G., Warden, P., Mattson, P. and Kanter, D., 2021. Data Engineering for Everyone. *arXiv preprint arXiv:2102.11447*.
300. Renz, A. and Vladova, G., 2021. Reinvigorating the Discourse on Human-Centered Artificial Intelligence in Educational Technologies. *Technology Innovation Management Review*, 11(5).
301. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
302. Richards NM, King JH (2013) Three paradoxes of big data. *Stan L Rev Online* 66:41
303. Ritter, S., Yudelso, M., Fancsali, S. E., & Berman, S. R. (2016). How Mastery Learning Works at Scale. *Proceedings of the Third (2016) ACM*

- Conference on Learning @ Scale*, 71–79.  
<https://doi.org/10.1145/2876034.2876039>
304. Romero, C. and Ventura, S., 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), pp.12-27.
305. Roos Jr, L.L., Nicol, J.P. and Cageorge, S.M., 1987. Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of chronic diseases*, 40(1), pp.41-49.
306. Roscher, R., Bohn, B., Duarte, M.F. and Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8, pp.42200-42216.
307. Roselli, D., Matthews, J. and Talagala, N., 2019, May. Managing bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 539-544).
308. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp.206-215.
309. Russell, S., Dewey, D. and Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, [online] 36(4), p.105. Available at:  
<https://www.aaai.org/ojs/index.php/aimagazine/article/view/2577>.
310. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
311. Samek, W, Wiegand, T, Müller, K-R. August 2017; Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv e-prints. arXiv:1708.08296.
312. Sapsford, R. and Jupp, V. eds., 1996. *Data collection and analysis*. Sage.
313. Sari, I.P., Al-Khowarizmi, A.K. and Batubara, I.H., 2021. Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process. *Journal of*



*Computer Science, Information Technology and Telecommunication Engineering*, 2(1), pp.139-144.

314. Satariano, A. (2020). British Grading Debacle Shows Pitfalls of Automating Government. *The New York Times*. [online] 20 Aug. Available at: <https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html>.
315. Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35- 37). Windsor, England: NFER-NELSON.
316. Schwarzer, R., Diehl, M., & Schmitz, G. S. (1999). Self-regulation. *Berlin: Freie Universtitat Berlin. Pridobljeno*, 20(3), 2007.
317. Scott, C & Sutton, R. (2009). Emotions and Change During Professional Development for Teachers: A Mixed Method Study. *Journal of Mixed Methods Research*, 3(2), 151-171.
318. Sellgren, K. 2018. Exam Boards Police Social Media in Cheating Crackdown. *British Broadcasting Corporation (BBC)*. 26 Jul 2018. Accessed 25 May 2020. <https://www.bbc.com/news/education-44965465>.
319. Senn, S., 2003. Bayesian, likelihood, and frequentist approaches to statistics. *Applied Clinical Trials*, 12(8), pp.35-38.
320. Setiono, R., Leow, W.K. and Thong, J., 2000. Opening the neural network black box: an algorithm for extracting rules from function approximating artificial neural networks. *ICIS 2000 Proceedings*, p.17.
321. Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R. and Khovanova, N., 2019. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, 52, pp.456-462.
322. Shawky, D. and Badawi, A., 2018, February. A reinforcement learning-based adaptive learning system. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 221-231). Springer, Cham.
323. Shoshan, A., Bhonker, N., Kviatkovsky, I. and Medioni, G., 2021. GAN-control: Explicitly controllable GANs. *arXiv preprint arXiv:2101.02477*.

324. Shreyas, V., Bharadwaj, S.N., Srinidhi, S., Ankith, K.U. and Rajendra, A.B., 2020. Self-driving cars: An overview of various autonomous driving systems. *Advances in Data and Information Sciences*, pp.361-371.
325. Siau, K. and Wang, W., 2020. Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, 31(2), pp.74-87.
326. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. and Lillicrap, T., 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), pp.1140-1144.
327. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. and Chen, Y., 2017. Mastering the game of go without human knowledge. *nature*, 550(7676), pp.354-359.
328. Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). Ieee.
329. Singla, K. and Biswas, S., 2021, January. Machine learning explainability method for the multi-label classification model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)* (pp. 337-340). IEEE.
330. Skitka, L.J., Mosier, K. and Burdick, M.D., 2000. Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4), pp.701-717.
331. Skrbinjek, V. and Dermol, V., 2019. Predicting students' satisfaction using a decision tree. *Tertiary Education and Management*, 25(2), pp.101-113.
332. Smith, E., 2008. Pitfalls and promises: The use of secondary data analysis in educational research. *British Journal of Educational Studies*, 56(3), pp.323-339.
333. Smith, H., 2020. Algorithmic bias: should students pay the price?. *Ai & Society*, 35(4), pp.1077-1078.

334. Spinner, T., Schlegel, U., Schäfer, H. and El-Assady, M., 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1), pp.1064-1074.
335. Stathoulopoulos, K. and Mateos-Garcia, J.C., 2019. Gender diversity in AI research. Available at SSRN 3428240.
336. Stohl C, Stohl M and Leonardi PM (2016) Managing opacity: information visibility and the paradox of transparency in the digital age. *International Journal of Communication Systems* 10: 123–137.
337. Stohl C, Stohl M and Leonardi PM (2016) Managing opacity: information visibility and the paradox of transparency in the digital age. *International Journal of Communication Systems* 10: 123–137.
338. Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S. and Müller, K.R., 2021. Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2), pp.392-413.
339. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H. and Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1199-1208).
340. Suresh, H. and Gutttag, J. (2019). *A Framework for Understanding Unintended Consequences of Machine Learning*. [online] . Available at: <https://arxiv.org/pdf/1901.10002.pdf>.
341. Surfshark. (2021). *Facial Recognition Map*. [online] Available at: <https://surfshark.com/facial-recognition-map>.
342. Sutton, R.S. and Barto, A.G., 1998. *Introduction to reinforcement learning* (Vol. 135). Cambridge: MIT press.
343. Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT press.
344. Tae, K.H., Roh, Y., Oh, Y.H., Kim, H. and Whang, S.E., 2019, June. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning* (pp. 1-4).

345. Tahiru, F., 2021. AI in Education: A Systematic Literature Review. *Journal of Cases on Information Technology (JCIT)*, 23(1), pp.1-20.
346. Tamboli, A., 2019. Good AI in the Hands of Bad Users. In *Keeping Your AI Under Control* (pp. 55-65). Apress, Berkeley, CA.
347. Tan, S., Caruana, R., Hooker, G. and Lou, Y., 2017. Detecting bias in black-box models using transparent model distillation. arXiv preprint arXiv:1710.06169.
348. Thammasiri, D., Delen, D., Meesad, P. and Kasap, N., 2014. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2), pp.321-330.
349. The Economist. (2018). *Why Uber's self-driving car killed a pedestrian*. [online] Available at: <https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian>.
350. The Institute for Ethical AI in Education, The Ethical Framework for AI in Education. (2021). [online] . Available at: <https://fb77c667c4d6e21c1e06.b-cdn.net/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf>
351. Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), 230–241. <https://doi.org/10.1080/09540091.2017.1310182>
352. Thomas, P.S., Castro, B., Barto, A.G., Giguere, S., Yuriy Brun and Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, [online] 366(6468), pp.999–1004. Available at: <https://science.sciencemag.org/content/366/6468/999>.
353. Thomas, S.L. and Heck, R.H., 2001. Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in higher education*, 42(5), pp.517-540.
354. Tondeur, J., van Braak, J., Siddiq, F. and Scherer, R., 2016. Time for a new approach to prepare future teachers for educational technology use: Its meaning and measurement. *Computers & Education*, 94, pp.134-150.

355. Tsymbal, A., 2004. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin, 106(2)*, p.58.
356. Turilli, M. and Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11(2)*, pp.105–112.
357. Tzeng, F.Y. and Ma, K.L., 2005. *Opening the black box-data driven visualization of neural networks* (pp. 383-390). IEEE.
358. Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.L.A., Elkhatib, Y., Hussain, A. and Al-Fuqaha, A., 2019. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access, 7*, pp.65579-65615.
359. Valera, J., J. Valera, and Y. Gelogo. 2015. “A Review on Facial Recognition for Online Learning Authentication.” In 8th International Conference on Bio-Science and Bio-Technology (BSBT), 16–19. IEEE.
360. Vandamme, J.P., Meskens, N. and Superby, J.F., 2007. Predicting academic performance by data mining methods. *Education Economics, 15(4)*, p.405.
361. Veale, M. and Binns, R., 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4(2)*, p.2053951717743530.
362. Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare '18: Proceedings of the International Workshop on Software Fairness, 1–7. <https://doi.org/10.1145/3194770.3194776>
363. Vigil, A., 2016. *Building explainable random forest models with applications in protein functional analysis* (Doctoral dissertation, San Francisco State University).
364. Vincent, J. (2018). *Google “fixed” its racist algorithm by removing gorillas from its image-labeling tech*. [online] The Verge. Available at: <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>.
365. Vinothkanna, M.R., 2020. A Survey on Novel Estimation Approach of Motion Controllers for Self-Driving Cars. *Journal of Electronics, 2(04)*, pp.211-219.
366. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P. and Oh, J.,

2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), pp.350-354.
367. W. Dieterich, C. Mendoza, and T. Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity, 2016
368. Wagenmakers, E.J., Lee, M., Lodewyckx, T. and Iverson, G.J., 2008. Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181-207). Springer, New York, NY.
369. Waheed, H., Anas, M., Hassan, S.U., Aljohani, N.R., Alelyani, S., Edifor, E.E. and Nawaz, R., 2021. Balancing sequential data to predict students at-risk using adversarial networks. *Computers & Electrical Engineering*, 93, p.107274.
370. Wang, A., Wan, G., Cheng, Z. and Li, S., 2009, November. An incremental extremely random forest classifier for online learning and tracking. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (pp. 1449-1452). IEEE.
371. Wang, H., Wang, L., & Liu, C. (2018). Employee Competitive Attitude and Competitive Behavior Promote Job-Crafting and Performance: A Two-Component Dynamic Model. *Frontiers in psychology*
372. Wang, Y. and Yao, Q., 2019. Few-shot learning: A survey.
373. Waters, A., & Miikkulainen, R. (2014). GRADE: Machine Learning Support for Graduate Admissions. *AI Magazine*, 35(1), 64.  
<https://doi.org/10.1609/aimag.v35i1.2504>
374. Watters, A. (2015). Education Technology and Skinner's Box. Retrieved from <http://hackededucation.com/2015/02/10/skinners-box>
375. Watters, A. (2017). Dunces' App: How Silicon Valley's brand of behaviorism has entered the classroom. *The Baffler*. Retrieved from <https://thebaffler.com/latest/behaviorismeducation-watters>
376. Webb, G.I., Hyde, R., Cao, H., Nguyen, H.L. and Petitjean, F., 2016. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4), pp.964-994.
377. Weller, A. (2017). Challenges for transparency. arXiv preprint [arXiv:1708.01870](https://arxiv.org/abs/1708.01870).
378. West, S.M., Whittaker, M. and Crawford, K., 2019. Discriminating systems: Gender, race and power in AI. *AI now Institute*, pp.1-33.

379. West, J. 2017. "Data, Democracy and School Accountability: Controversy Over School Evaluation in the Case of DeVasco High School." *Big Data and Society* 4 (1): 1–16. <https://journals.sagepub.com/doi/full/10.1177/2053951717702408>.
380. Whittlestone, J., Nyrup, R., Alexandrova, A. and Cave, S., 2019, January. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195-200).
381. Wilson, B.G., 1997. Thoughts on theory in educational technology. *Educational Technology*, 37(1), pp.22-27.
382. Wilson, C., & Scott, B. (2017). Adaptive systems in education: a review and conceptual unification. *The International Journal of Information and Learning Technology*, 34(1), 2-19. doi:http://dx.doi.org/10.1108/IJILT-09-2016-0040
383. Wischmeyer, T., 2020. Artificial intelligence and transparency: opening the black box. In *Regulating artificial intelligence* (pp. 75-101). Springer, Cham.
384. Wise, A.F., Speer, J., Marbouti, F. and Hsiao, Y.T., 2013. Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors. *Instructional Science*, 41(2), pp.323-343.
385. Wolff, A., Z. Zdrahal, A. Nikolov, and M. Pantucek. 2013. Improving Retention: Predicting At-Risk Students by Analysing Clicking Behaviour in a Virtual Learning Environment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 145–149). ACM.
386. Wu, M., Goodman, N., Piech, C. and Finn, C., 2021. ProtoTransformer: A Meta-Learning Approach to Providing Student Feedback. *arXiv preprint arXiv:2107.14035*.
387. Wu, M., Mosse, M., Goodman, N. and Piech, C., 2019, July. Zero shot learning for code education: Rubric sampling with deep learning inference. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 782-790).

388. Wu, X., Zhu, X., Wu, G.Q. and Ding, W., 2013. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1), pp.97-107.
389. Xu, Y. and Wilson, K., 2021, April. Early Alert Systems During a Pandemic: A Simulation Study on the Impact of Concept Drift. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 504-510).
390. Yampolskiy, R.V. and Spellchecker, M.S., 2016. Artificial intelligence safety and cybersecurity: A timeline of AI failures. *arXiv preprint arXiv:1610.07997*.
391. Yan, S., Kao, H.T. and Ferrara, E., 2020, October. Fair class balancing: enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1715-1724).
392. Yang, Y., Kandogan, E., Li, Y., Sen, P. and Lasecki, W., 2019. A Study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. In *IUI Workshops*
393. Yanisky-Ravid, S. and Hallisey, S.K., 2019. Equality and Privacy by Design: A New Model of Artificial Intelligence Data Transparency Via Auditing, Certification, and Safe Harbor Regimes. *Fordham Urb. LJ*, 46, p.428.
394. Yin, M., Wortman Vaughan, J. and Wallach, H., 2019, May. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-12).
395. Yu, K.H., Beam, A.L. and Kohane, I.S., 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10), pp.719-731.
396. Yudelson, M. V., Fancsali, S. E., Ritter, S., Berman, S. R., Nixon, T., & Joshi, A. (2014). Better Data Beat Big Data. *Proceedings of the 7th International Conference on Educational Data Mining*, 205–208.
397. Zajac, M., 2009. Using learning styles to personalize online learning. *Campus-wide information systems*.



398. Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Sci Technol Human Values* 41(1):118–132
399. Završnik, A., 2021. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18(5), pp.623-642.
400. Zetzsche, D.A., Arner, D.W., Buckley, R.P. and Tang, B., 2020. Artificial Intelligence in Finance: Putting the Human in the Loop. *CFTE Academic Paper Series: Centre for Finance, Technology and Entrepreneurship*, (1).
401. Zhang, M., Cheng, X., Copeland, D., Desai, A., Guan, M.Y., Brat, G.A. and Yeung, S., 2020. Using Computer Vision to Automate Hand Detection and Tracking of Surgeon Movements in Videos of Open Surgery. In *AMIA Annual Symposium Proceedings* (Vol. 2020, p. 1373). American Medical Informatics Association.
402. Zhang, Y., An, R., Cui, J. and Shang, X., 2021, April. Undergraduate grade prediction in chinese higher education using convolutional neural networks. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 462-468).
403. Zhang, Z., Beck, M.W., Winkler, D.A., Huang, B., Sibanda, W. and Goyal, H., 2018. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11).
404. Zhao, S., 2021, October. Facial Recognition in Educational Context. In *2021 International Conference on Public Relations and Social Sciences (ICPRSS 2021)* (pp. 10-17). Atlantis Press.
405. Zhao, Y. and Liu, G., 2018. How do teachers face educational changes in artificial intelligence era. *Advances in Social Science, Education and Humanities Research*, 300, pp.47-50.
406. Zhou, S.M., Lyons, R.A., Bodger, O., Demmler, J.C. and Atkinson, M.D., 2010, July. SVM with entropy regularization and particle swarm optimization for identifying children's health and socioeconomic determinants of education attainments using linked datasets. In *The 2010*

- International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8).  
IEEE.
407. Zhou, Y., Kantarcioglu, M. and Clifton, C., 2021. Improving Fairness of AI Systems with Lossless De-biasing. *arXiv preprint arXiv:2105.04534*.
408. Zierau, N., Flock, K., Janson, A., Söllner, M. and Leimeister, J.M., 2021. The Influence of AI-Based Chatbots and Their Design on Users' Trust and Information Sharing in Online Loan Applications.
409. Žliobaitė, I. and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2), pp.183–201.
410. Žliobaitė, I., 2010. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*.

# Appendices

## Appendix 1: The details of different features used in personality surveys

### Personality

The detailed description of the dimensions used in the personality survey were from “Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>” is given below:

Five dimensions: categories (1 = very low, 2=low, 3 = average, 4 = high, 5=very high)

- **Neuroticism:** Freud originally used the term *neurosis* to describe a condition marked by mental distress, emotional suffering, and an inability to cope effectively with the normal demands of life. He suggested that everyone shows some signs of neurosis, but that we differ in our degree of suffering and our specific symptoms of distress. Today neuroticism refers to the tendency to experience negative feelings. Those who score high on Neuroticism may experience primarily one specific negative feeling such as anxiety, anger, or depression, but are likely to experience several of these emotions. People high in neuroticism are emotionally reactive. They respond emotionally to events that would not affect most people, and their reactions tend to be more intense than normal. They are more likely to interpret ordinary situations as threatening, and minor frustrations as hopelessly difficult. Their negative emotional reactions tend to persist for unusually long periods of time, which means they are often in a bad mood. These problems in emotional regulation can diminish a neurotic's ability to think clearly, make decisions, and cope effectively with stress. At the other end of the scale, individuals who score low in neuroticism are less easily upset and are less emotionally reactive.

They tend to be calm, emotionally stable, and free from persistent negative feelings. Freedom from negative feelings does not mean that low scorers experience a lot of positive feelings; frequency of positive emotions is a component of the Extraversion domain.

- **Extraversion:** Extraversion is marked by pronounced engagement with the external world. Extraverts enjoy being with people, are full of energy, and often experience positive emotions. They tend to be enthusiastic, action-oriented, individuals who are likely to say "Yes!" or "Let's go!" to opportunities for excitement. In groups they like to talk, assert themselves, and draw attention to themselves. Introverts lack the exuberance, energy, and activity levels of extraverts. They tend to be quiet, low-key, deliberate, and disengaged from the social world. Their lack of social involvement should not be interpreted as shyness or depression; the introvert simply needs less stimulation than an extravert and prefers to be alone. The independence and reserve of the introvert is sometimes mistaken as unfriendliness or arrogance. In reality, an introvert who scores high on the agreeableness dimension will not seek others out but will be quite pleasant when approached.
- **Openness to Experience:** Openness to Experience describes a dimension of cognitive style that distinguishes imaginative, creative people from down-to-earth, conventional people. Open people are intellectually curious, appreciative of art, and sensitive to beauty. They tend to be, compared to closed people, more aware of their feelings. They tend to think and act in individualistic and nonconforming ways. Intellectuals typically score high on Openness to Experience; consequently, this factor has also been called *Culture* or *Intellect*. Nonetheless, Intellect is probably best regarded as one aspect of openness to experience. Scores on Openness to Experience are only modestly related to years of education and scores on standard intelligent tests. Another characteristic of the open cognitive style is a facility for thinking in symbols and abstractions far removed from concrete experience. Depending on the individual's specific intellectual abilities, this symbolic cognition may take the form of mathematical,

logical, or geometric thinking, artistic and metaphorical use of language, music composition or performance, or one of the many visual or performing arts. People with low scores on openness to experience tend to have narrow, common interests. They prefer the plain, straightforward, and obvious over the complex, ambiguous, and subtle. They may regard the arts and sciences with suspicion, regarding these endeavours as abstruse or of no practical use. Closed people prefer familiarity over novelty; they are conservative and resistant to change. Openness is often presented as healthier or more mature by psychologists, who are often themselves open to experience. However, open and closed styles of thinking are useful in different environments. The intellectual style of the open person may serve a professor well, but research has shown that closed thinking is related to superior job performance in police work, sales, and a number of service occupations.

- **Agreeableness:** Agreeableness reflects individual differences in concern with cooperation and social harmony. Agreeable individuals value getting along with others. They are therefore considerate, friendly, generous, helpful, and willing to compromise their interests with others'. Agreeable people also have an optimistic view of human nature. They believe people are basically honest, decent, and trustworthy. Disagreeable individuals place self-interest above getting along with others. They are generally unconcerned with others' well-being, and therefore are unlikely to extend themselves for other people. Sometimes their skepticism about others' motives causes them to be suspicious, unfriendly, and uncooperative. Agreeableness is obviously advantageous for attaining and maintaining popularity. Agreeable people are better liked than disagreeable people. On the other hand, agreeableness is not useful in situations that require tough or absolute objective decisions. Disagreeable people can make excellent scientists, critics, or soldiers.
- **Conscientiousness :** Conscientiousness concerns the way in which we control, regulate, and direct our impulses. Impulses are not

inherently bad; occasionally time constraints require a snap decision, and acting on our first impulse can be an effective response. Also, in times of play rather than work, acting spontaneously and impulsively can be fun. Impulsive individuals can be seen by others as colorful, fun-to-be-with, and zany. Nonetheless, acting on impulse can lead to trouble in a number of ways. Some impulses are antisocial. Uncontrolled antisocial acts not only harm other members of society, but also can result in retribution toward the perpetrator of such impulsive acts. Another problem with impulsive acts is that they often produce immediate rewards but undesirable, long-term consequences. Examples include excessive socializing that leads to being fired from one's job, hurling an insult that causes the breakup of an important relationship, or using pleasure-inducing drugs that eventually destroy one's health. Impulsive behaviour, even when not seriously destructive, diminishes a person's effectiveness in significant ways. Acting impulsively disallows contemplating alternative courses of action, some of which would have been wiser than the impulsive choice. Impulsivity also sidetracks people during projects that require organized sequences of steps or stages. Accomplishments of an impulsive person are therefore small, scattered, and inconsistent. A hallmark of intelligence, what potentially separates human beings from earlier life forms, is the ability to think about future consequences before acting on an impulse. Intelligent activity involves contemplation of long-range goals, organizing and planning routes to these goals, and persisting toward one's goals in the face of short-lived impulses to the contrary. The idea that intelligence involves impulse control is nicely captured by the term prudence, an alternative label for the Conscientiousness domain. Prudent means both wise and cautious. Persons who score high on the Conscientiousness scale are, in fact, perceived by others as intelligent. The benefits of high conscientiousness are obvious. Conscientious individuals avoid trouble and achieve high levels of success through purposeful planning and persistence. They are also positively regarded by others as intelligent and reliable. On the negative side, they can be compulsive perfectionists and workaholics.

Furthermore, extremely conscientious individuals might be regarded as stuffy and boring. Unconscientious people may be criticized for their unreliability, lack of ambition, and failure to stay within the lines, but they will experience many short-lived pleasures and they will never be called stuffy.

#### **Sub dimensions (levels 1-10)**

- **Neuroticism:**
  - **Depression.** This scale measures the tendency to feel sad, dejected, and discouraged. High scorers lack energy and have difficulty initiating activities. Low scorers tend to be free from these depressive feelings. Your level of depression is average.
  
- **Extraversion:**
  - **Excitement-Seeking.** High scorers on this scale are easily bored without high levels of stimulation. They love bright lights and hustle and bustle. They are likely to take risks and seek thrills. Low scorers are overwhelmed by noise and commotion and are adverse to thrill-seeking. Your level of excitement-seeking is low.
  
- **Openness to Experience:**
  - **Imagination.** To imaginative individuals, the real world is often too plain and ordinary. High scorers on this scale use fantasy as a way of creating a richer, more interesting world. Low scorers are on this scale are more oriented to facts than fantasy. Your level of imagination is low.
  
  - **Artistic Interests.** High scorers on this scale love beauty, both in art and in nature. They become easily involved and absorbed in artistic and natural events. They are not necessarily artistically trained nor talented, although many will be. The defining features of this scale are *interest in*, and *appreciation of* natural and artificial beauty. Low scorers lack aesthetic sensitivity and interest in the arts. Your level of artistic interests is low.

- **Intellect.** Intellect and artistic interests are the two most important, central aspects of openness to experience. High scorers on Intellect love to play with ideas. They are open-minded to new and unusual ideas, and like to debate intellectual issues. They enjoy riddles, puzzles, and brain teasers. Low scorers on Intellect prefer dealing with either people or things rather than ideas. They regard intellectual exercises as a waste of time. Intellect should not be equated with intelligence. Intellect is an intellectual style, not an intellectual ability, although high scorers on Intellect score slightly higher than low-Intellect individuals on standardized intelligence tests. Your level of intellect is average.
- **Agreeableness:**
  - **Trust.** A person with high trust assumes that most people are fair, honest, and have good intentions. Persons low in trust see others as selfish, devious, and potentially dangerous. Your level of trust is low.
  - **Altruism.** Altruistic people find helping other people genuinely rewarding. Consequently, they are generally willing to assist those who are in need. Altruistic people find that doing things for others is a form of self-fulfillment rather than self-sacrifice. Low scorers on this scale do not particularly like helping those in need. Requests for help feel like an imposition rather than an opportunity for self-fulfillment. Your level of altruism is low.
- **Conscientiousness :**
  - **Orderliness.** Persons with high scores on orderliness are well-organized. They like to live according to routines and schedules. They keep lists and make plans. Low scorers tend to be disorganized and scattered. Your level of orderliness is average.
  - **Dutifulness.** This scale reflects the strength of a person's sense of duty and obligation. Those who score high on this scale have a strong sense of moral obligation. Low scorers find contracts,



rules, and regulations overly confining. They are likely to be seen as unreliable or even irresponsible. Your level of dutifulness is low.

- **Cautiousness.** Cautiousness describes the disposition to think through possibilities before acting. High scorers on the Cautiousness scale take their time when making decisions. Low scorers often say or do first thing that comes to mind without deliberating alternatives and the probable consequences of those alternatives. Your level of cautiousness is high.

Emotion Regulation score (The higher the scores the greater the use of the emotion regulation strategy) from "Gross, J. J., & John, O. P. (2003). Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being. *Journal of Personality and Social Psychology*, 85(2), 348–362. <https://doi.org/10.1037/0022-3514.85.2.348>"

- **Reappraisal score:** Cognitive reappraisal is a form of cognitive change that involves construing a potentially emotion-eliciting situation in a way that changes its emotional impact (Lazarus & Alfert, 1964). For example, during an admissions interview, one might view this as an opportunity to find out how much one likes the school, rather than as a test of one's worth.
- **Suppression score:** Expressive suppression is a form of response modulation that involves inhibiting ongoing emotion-expressive behaviour (Gross, 1998). For example, one might keep a poker face while holding a great hand during a card game.

Cognitive Reflection Test (CRT) from "Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>"

- **System 2 score:** Ability to override intuitive System 1 responses and engage in analytic thinking, which leads to reduced biases and more normative responses, extended delay of gratification and normative

responses to risky choice, disposition effect, individual reliance on the System 1.

Adult - Decision Making Competence (ADMC) from “De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92(5), 938–956. <https://doi.org/10.1037/0022-3514.92.5.938>”

- **Sunk cost** - Resistance to Sunk Costs
- **Confidence** - Under/Overconfidence, and confidence accuracy score
- **Risk perception** - Consistency in Risk Perception

General Self-Efficacy Scale (GSE) from “Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user’s portfolio. Causal and control beliefs (pp. 35- 37). Windsor, England: NFER-NELSON.”

- **General Self efficacy** - Self-efficacy expectancies refer to personal action control or agency. A person who believes in being able to cause an event can conduct a more active and self-determined life course. This “can do” cognition mirrors a sense of control over one’s environment. It reflects the belief of being able to control challenging environmental demands by means of taking adaptive action. It can be regarded as a self-confident view of one’s capability to deal with certain life stressors. The General Self-Efficacy Scale is correlated to emotion, optimism, work satisfaction. Negative coefficients were found for depression, stress, health complaints, burnout, and anxiety.

Self-Regulation

Schwarzer, R., Diehl, M., & Schmitz, G. S. (1999). Self-regulation. *Berlin: Freie Universtitat Berlin. Pridobljeno*, 20(3), 2007.

- **Self regulation score:** when individuals are in the phase of goal-pursuit, and face difficulties in maintaining their action. In such a maintenance situation it is required to focus attention on the task at

hand and to keep a favourable emotional balance. Thus, attention-regulation and emotion-regulation are reflected in these scale items. The Proactive Coping Inventory (PCI) from "Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., & Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14)."

- **Proactive coping** - combines autonomous goal setting with self-regulatory goal attainment cognitions and behaviour.
- **Preventive coping** - deals with anticipation of potential stressors and the initiation of preparation before these stressors develop fully. Preventive coping is distinct from proactive coping. Preventive coping effort refers to a potential threat in future by considering experience, anticipation or knowledge. In comparison, proactive coping is not based on threat but is driven by goal striving.
- **Instrumental support seeking** - focuses on obtaining advice, information and feedback from people in one's social network when dealing with stressors.
- **Reflective coping** - describes simulation and contemplation about a variety of possible behavioural alternatives by comparing their imagined effectiveness and includes brainstorming, analysing problems and resources, and generating hypothetical plans of action.
- **Strategic planning** - focuses on the process of generating a goal-oriented schedule of action in which extensive tasks are broken down into manageable components.
- **Avoidance coping** - eludes action in a demanding situation by delaying

**Trait Competitiveness** from "Brown, S. P., Cron, W. L., & Slocum Jr, J. W. (1998). Effects of trait competitiveness and perceived intraorganizational competition on salesperson goal setting and performance. *Journal of Marketing*, 62(4), 88-98."

- Trait competitiveness is conceptualized as an aspect of personality that involves "the enjoyment of interpersonal competition and the desire to win and be better than others".

**Competitive attitude scale** from “Wang, H., Wang, L., & Liu, C. (2018). Employee Competitive Attitude and Competitive Behavior Promote Job-Crafting and Performance: A Two-Component Dynamic Model. *Frontiers in psychology*”

- Competitive attitude – a belief concerning whether an individual likes competition.

**Competitive behaviour scale also from** “Wang, H., Wang, L., & Liu, C. (2018). Employee Competitive Attitude and Competitive Behaviour Promote Job-Crafting and Performance: A Two-Component Dynamic Model. *Frontiers in psychology*”

- Competitive behaviour - the actual actions people take or are inclined to take in a specific job or life environment to compete for resources or succeed over others.

Table 30: Description of the features used and academic references for each feature for training the recruitment tool’s models for an ed-tech company in financial services

	<b>Feature</b>	<b>Reference</b>
1	<b>Depression:</b> High scorers lack energy and have difficulty initiating activities.	Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality</i> , 51, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a>
2	<b>Excitement-Seeking:</b> High scorers are easily bored without high levels of stimulation.	Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality</i> , 51, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a>

3	<p><b>Imagination:</b> To imaginative individuals, the real world is often too plain and ordinary.</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>
4	<p><b>Artistic Interests:</b> High scorers on this scale love beauty, both in art and in nature.</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>
5	<p><b>Intellect:</b> High scorers on Intellect love to play with ideas.</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>
6	<p><b>Trust:</b> high trust assumes that most people are fair, honest, and have good intentions.</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>
7	<p><b>Altruism:</b> Altruistic people find helping other people genuinely rewarding.</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>
8	<p><b>Orderliness:</b> High are well-organized. They like to live by routines and schedules</p>	<p>Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality, 51</i>, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a></p>

9	<b>Dutifulness:</b> reflects the strength of a person's sense of duty and obligation.	Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality</i> , 51, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a>
10	<b>Cautiousness:</b> thinking through possibilities before acting.	Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. <i>Journal of Research in Personality</i> , 51, 78–89. <a href="https://doi.org/10.1016/j.jrp.2014.05.003">https://doi.org/10.1016/j.jrp.2014.05.003</a>
11	<b>Reappraisal:</b> finding a way to find an upside to a difficult situation.	Gross, J. J., & John, O. P. (2003). Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being. <i>Journal of Personality and Social Psychology</i> , 85(2), 348–362. <a href="https://doi.org/10.1037/0022-3514.85.2.348">https://doi.org/10.1037/0022-3514.85.2.348</a>
12	<b>Suppression score:</b> inhibiting emotion-expressive behaviour, e.g. keeping a poker face.	Gross, J. J., & John, O. P. (2003). Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being. <i>Journal of Personality and Social Psychology</i> , 85(2), 348–362. <a href="https://doi.org/10.1037/0022-3514.85.2.348">https://doi.org/10.1037/0022-3514.85.2.348</a>
13	<b>System 2:</b> Ability to override intuitive responses and engage in analytic thinking,	Frederick, S. (2005). Cognitive Reflection and Decision Making. <i>Journal of Economic Perspectives</i> , 19(4), 25–42. <a href="https://doi.org/10.1257/089533005775196732">https://doi.org/10.1257/089533005775196732</a>
14	<b>Sunk cost:</b> Resistance to basing decisions on the	De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. <i>Journal of Personality and</i>

	amount of money already committed.	<i>Social Psychology</i> , 92(5), 938–956. <a href="https://doi.org/10.1037/0022-3514.92.5.938">https://doi.org/10.1037/0022-3514.92.5.938</a>
15	<b>Confidence:</b> The feeling associated with the expectation of one’s ability to successfully carry out a task	De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. <i>Journal of Personality and Social Psychology</i> , 92(5), 938–956. <a href="https://doi.org/10.1037/0022-3514.92.5.938">https://doi.org/10.1037/0022-3514.92.5.938</a>
16	<b>Risk perception:</b> Consistency in Risk Perception	De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. <i>Journal of Personality and Social Psychology</i> , 92(5), 938–956. <a href="https://doi.org/10.1037/0022-3514.92.5.938">https://doi.org/10.1037/0022-3514.92.5.938</a>
17	<b>General Self efficacy:</b> belief in one’s ability to successfully carry out a task.	Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user’s portfolio. Causal and control beliefs (pp. 35- 37). Windsor, England: NFER-NELSON.
18	<b>Self-regulation:</b> Maintaining calm and focus in the face of difficulties.	Schwarzer, R., Diehl, M., & Schmitz, G. S. (1999). Self-regulation. <i>Berlin: Freie Universtitat Berlin. Pridobljeno</i> , 20(3), 2007.]
19	<b>Proactive coping:</b> combines autonomous goal setting with self-regulatory goal attainment cognitions and behaviour.	Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., & Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).

20	<p><b>Preventive coping:</b> Imagining possible future threats and ways to deal with them.</p>	<p>Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., &amp; Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).</p>
21	<p><b>Instrumental support seeking:</b> obtaining advice, information and feedback when dealing with stressors.</p>	<p>Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., &amp; Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).</p>
22	<p><b>Reflective coping:</b> Imagining possible future courses of action and their outcomes.</p>	<p>Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., &amp; Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).</p>
23	<p><b>Strategic planning:</b> generating a goal-oriented schedule of action in which tasks are broken down into sub-tasks.</p>	<p>Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., &amp; Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).</p>
24	<p><b>Trait competitiveness:</b> the enjoyment of interpersonal competition</p>	<p>Brown, S. P., Cron, W. L., &amp; Slocum Jr, J. W. (1998). Effects of trait competitiveness and perceived intraorganizational competition on</p>



	and the desire to win and be better than others.	salesperson goal setting and performance. <i>Journal of Marketing</i> , 62(4), 88-98.]
25	<b>Competitive attitude:</b> a belief concerning whether an individual likes competition.	Wang, H., Wang, L., & Liu, C. (2018). Employee Competitive Attitude and Competitive Behavior Promote Job-Crafting and Performance: A Two-Component Dynamic Model. <i>Frontiers in psychology</i> , 9.]
26	<b>Competitive behaviour:</b> the actual actions people take or are inclined to take in order to succeed.	Wang, H., Wang, L., & Liu, C. (2018). Employee Competitive Attitude and Competitive Behavior Promote Job-Crafting and Performance: A Two-Component Dynamic Model. <i>Frontiers in psychology</i> , 9.]
27	<b>Adult Decision Making Confidence Accuracy Score</b>	De Bruin, W. B., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. <i>Journal of Personality and Social Psychology</i> , 92(5), 938–956. <a href="https://doi.org/10.1037/0022-3514.92.5.938">https://doi.org/10.1037/0022-3514.92.5.938</a>
28	<b>Sum Avoidance Coping</b> OR Sum Avoidance Coping Normalised	Greenglass, E., Schwarzer, R., Jakubiec, D., Fiksenbaum, L., & Taubert, S. (1999, July). The proactive coping inventory (PCI): A multidimensional research instrument. In 20th International Conference of the Stress and Anxiety Research Society (STAR), Cracow, Poland (Vol. 12, p. 14).

## Appendix 2: The Survey's English version of the reduced set of items

The survey used for traders whose first language was not English is given below:

### Questionnaire 1

#### Instructions:

This survey presents true/false questions about various aspects of everyday life. Please indicate, for each statement, whether you believe it to be true or false, by selecting the "true" or "false". You may think that some items do not have a clear-cut answer. For those items, please try to give the answer that would be true in general, or in most cases.

Please read through the following examples to find out more about this survey.

#### Example 1:

##### Pittsburgh's hockey team is the Bruins.

We want you to do two things:

First, answer the question. In this example, you might think "No, it's the Penguins. So the statement is FALSE." Then you would choose 'False'.

**Pittsburgh's hockey team is the Bruins.** This statement is [True / False].

Second, think about how sure you are of your answer. Give a number from 50% to 100%. In other words, what is the percent chance that you are right? Choose one of the numbers on the scale.

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

If your answer is a total guess, circle 50%. This means that there is a 50% chance that you are right, and a 50% chance that you are wrong. If you are absolutely sure, circle 100%. If you aren't sure, then circle a number in between, to show how sure you are.

In this example, you might think "I'm absolutely sure it's false, so 100%." So you would choose 100%.

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

Please read the examples below. They show answers given by other people. Read them closely, and make sure you understand their answers.

**Example 2:**

**Thanksgiving Day is on the fourth Thursday of November.**

- Yes, I think that's when Thanksgiving is. I would say TRUE.
- I'm pretty sure, but it might be on the third Thursday of November, so 80%.

Your answer would look like this:

**Thanksgiving Day is on the fourth Thursday of November.** This statement is [ True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**Example 3:**

**Amman is the capital of Jordan.**

- I really don't know, so I'll just take a guess. I'll say, uh, TRUE.
- I'm guessing, so 50%.

Your answer would look like this:

**Amman is the capital of Jordan.** This statement is [ True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**Example 4:**

**The Hudson River doesn't run past New York City.**

- Oh yes it does! I think it's one of the rivers. So that's FALSE.
- I'm almost positive that's false, so I'll say 90%.

Your answer would look like this:

**The Hudson River doesn't run past New York City.** This statement is [ True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**Example 5:**

**Bill Clinton doesn't have a beard.**

- That's right, he doesn't. TRUE.
  - I think that's right, but I'm not sure, he might have grown one. I'll say 70%.
- Your answer would look like this:

**Bill Clinton doesn't have a beard. This statement is [ True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**That is the end of the practice. Please answer the following questions using the same way.**

**For each of the following statements, choose true or false to indicate your answer. Then choose a number on the scale to indicate how sure you are of your answer. The scale ranges from 50% (meaning that you were just guessing) to 100% (meaning that you were absolutely sure).**

- 1) Many smokers use the nicotine in cigarettes to treat depression.**  
This statement is [True / False]

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

- 2) Stress makes it easier to form bad habits.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

- 3) You can take wrinkles out of your clothes by putting them in the dryer with a damp towel.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**4) After a fight with your partner, you should not focus on who was to blame.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**5) There is no way to improve your memory.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**6) The grace period on your credit card is the amount of time you do not have to pay interest on outstanding payments.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**7) Red wine stains are easier to remove than beer stains.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**8) Muscles do not burn calories when you are at rest.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**9) Alcohol causes dehydration.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**10) Problems with in-laws contribute to more than 30% of divorces.**

This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**11) Homosexual couples are not legally allowed to adopt.** This

statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**12) A promotion means that you will get a more satisfying job.** This

statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**13) HMRC forms are available on-line.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**14) Procrastination is worse when you work in a cluttered environment.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
------------	------------	------------	------------	------------	-------------

Just guessing	...	Absolutely sure
---------------	-----	-----------------

**15) A venture capital fund invests in new businesses by providing startup capital.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**16) It is wise to handle all negotiations yourself, even if your opponent uses a lawyer.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**17) Carbohydrates are fattening no matter how much you eat of them.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**18) Young people face few stereotypes when looking for a job.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**19) It can be instructive for children to see their parents resolve a fight.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
------------	------------	------------	------------	------------	-------------

Just guessing	...	Absolutely sure
---------------	-----	-----------------

**20) There are non-profit organizations that help people with debt counselling.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**21) Assertive behaviour makes your brain experience an increase in pleasure.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**22) Credit card companies can offer lower payments if you can come up with a lump sum settlement.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**23) Contracting a sexually transmitted disease is not an automatic sign that your partner has had an affair.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**24) Some sexually transmitted diseases can cause infertility.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure



--	--	--

**25) Self-employed people pay the same amount of taxes as people who work for an employer. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**26) When buying a new home, there is little need to have it inspected before you buy it. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**27) Creating a routine is an important step in getting unpleasant work done. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**28) Once you have experienced an event, your memory of it cannot be changed. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**29) Meditation slows the heart rate. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**30) If you get into an auto accident, let the other person take the lead in handling the details. This statement is [True / False].**

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**31) There is no way you can negotiate a lower rate with a credit card company.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**32) Obesity increases your risk of type 2 diabetes.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**33) Talking about sex helps romantic relationships.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

**34) Hard evidence is lacking that acupuncture helps you to quit smoking.** This statement is [True / False].

50% (1)	60% (2)	70% (3)	80% (4)	90% (5)	100% (6)
Just guessing	...				Absolutely sure

## Questionnaire 2

### Instructions:

Each of the following problems presents a choice between two options. Each problem is presented with a scale ranging from 1 (representing one option) through 6 (representing the other option). For each item, please choose the number on the scale that best reflects your relative preference between the two options.

**Problem 1**

You are buying a gold ring on layaway (a system of paying a deposit to secure an article for later purchase) for someone special. It costs \$200 and you have already paid \$100 on it, so you owe another \$100. One day, you see in the paper that a new jewellery store is selling the same ring for only \$90 as a special sale, and you can pay for it using layaway. The new store is across the street from the old one. If you decide to get the ring from the new store, you will not be able to get your money back from the old store, but you would save \$10 overall.

Would you be more likely to continue paying at the old store or buy from the new store?

1	2	3	4	5	6
Most likely to continue paying at the old store	...				Most likely to buy from the new store

**Problem 2**

You enjoy playing tennis, but you really love bowling. You just became a member of a tennis club, and of a bowling club, both at the same time. The membership to your tennis club costs \$200 per year and the membership to your bowling club \$50 per year. During the first week of both memberships, you develop an elbow injury. It is painful to play either tennis or bowling. Your doctor tells you that the pain will continue for about a year.

Would you be more likely to play tennis or bowling in the next six months?

1	2	3	4	5	6
Most likely to play tennis	...				Most likely to play bowling

**Problem 3**

You have been looking forward to this year's Halloween party. You have the right cape, the right wig, and the right hat. All week, you have been trying to perfect the outfit by cutting out a large number of tiny stars to glue to the cape and the hat, and you still need to glue them on. On the day of Halloween, you decide that the outfit looks better without all these stars you have worked so hard on.

Would you be more likely to wear the stars or go without?

1	2	3	4	5	6
Most likely to wear stars	...				Most likely to not wear stars

**Problem 4**

After a large meal at a restaurant, you order a big dessert with chocolate and ice cream. After a few bites you find you are full and you would rather not eat any more of it.

Would you be more likely to eat more or to stop eating it?

1	2	3	4	5	6
Most likely to eat more	...				Most likely to stop eating

**Problem 5**

You are in a hotel room for one night and you have paid \$6.95 to watch a movie on pay TV. Then you discover that there is a movie you would much rather like to see on one of the free cable TV channels. You only have time to watch one of the two movies.

Would you be more likely to watch the movie on pay TV or on the free cable channel?

1	2	3	4	5	6
Most likely to watch pay TV	...				Most likely to watch free cable

**Problem 6**

You have been asked to give a toast at your friend's wedding. You have worked for hours on this one story about you and your friend taking drivers' education, but you still have some work to do on it. Then you realize that you could finish writing the speech faster if you start over and tell the funnier story about the dance lessons you took together.

Would you be more likely to finish the toast about driving or rewrite it to be about dancing?

1	2	3	4	5	6
Most likely to write about driving	...				Most likely to write about dancing

**Problem 7**

You decide to learn to play a musical instrument. After you buy an expensive cello, you find you are no longer interested. Your neighbour is moving and you are excited that she is leaving you her old guitar, for free. You'd like to learn how to play it.

Would you be more likely to practice the cello or the guitar?

1	2	3	4	5	6
---	---	---	---	---	---

Most likely to play cello	...	Most likely to play guitar
---------------------------	-----	----------------------------

**Problem 8**

You and your friend are at a movie theatre together. Both you and your friend are getting bored with the storyline. You'd hate to waste the money spent on the ticket, but you both feel that you would have a better time at the coffee shop next door. You could sneak out without other people noticing.

Would you be more likely to stay or to leave?

1	2	3	4	5	6
Most likely to stay	...				Most likely to leave

**Problem 9**

You and your friend have driven halfway to a resort. Both you and your friend feel sick. You both feel that you both would have a much better weekend at home. Your friend says it is "too bad" you already drove halfway, because you both would much rather spend the time at home. You agree.

Would you be more likely to drive on or turn back?

1	2	3	4	5	6
Most likely to drive on	...				Most likely to turn back

**Problem 10**

You are painting your bedroom with a sponge pattern in your favourite colour. It takes a long time to do. After you finish two of the four walls, you realize you would have preferred the solid colour instead of the sponge pattern. You have enough paint left over to redo the entire room in the solid colour. It would take you the same amount of time as finishing the sponge pattern on the two walls you have left.

Would you be more likely to finish the sponge pattern or to redo the room in the solid colour?

1	2	3	4	5	6
Most likely to finish sponge pattern	...				Most likely to redo with a solid colour

**Questionnaire 3**

**Instructions and Items**

We would like to ask you some questions about your emotional life, in particular, how you control (that is, regulate and manage) your emotions. The questions below involve two distinct aspects of your emotional life. One is your emotional experience, or what you feel like inside. The other is your

emotional expression, or how you show your emotions in the way you talk, gesture, or behave. Although some of the following questions may seem similar to one another, they differ in important ways. For each item, please answer using the following scale:

Strongly disagree (1) - Disagree (2) - More or less disagree (3) - Neutral (4) - More or less agree (5) – Agree (6) – Strongly agree (7)

1. \_\_\_\_ When I want to feel more *positive* emotion (such as joy or amusement), I *change what I'm thinking about*.
2. \_\_\_\_ I keep my emotions to myself.
3. \_\_\_\_ When I want to feel less *negative* emotion (such as sadness or anger), I *change what I'm thinking about*.
4. \_\_\_\_ When I am feeling *positive* emotions, I am careful not to express them.
5. \_\_\_\_ When I'm faced with a stressful situation, I make myself *think about it* in a way that helps me stay calm.
6. \_\_\_\_ I control my emotions by *not expressing them*.
7. \_\_\_\_ When I want to feel more *positive* emotion, I *change the way I'm thinking* about the situation.
8. \_\_\_\_ I control my emotions by *changing the way I think* about the situation I'm in.
9. \_\_\_\_ When I am feeling *negative* emotions, I make sure not to express them.
10. \_\_\_\_ When I want to feel less *negative* emotion, I *change the way I'm thinking* about the situation.

#### Questionnaire 4

Instructions: below are three items that vary in difficulty. Answer as many as you can.

(1) A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? \_\_\_\_\_ (cents)

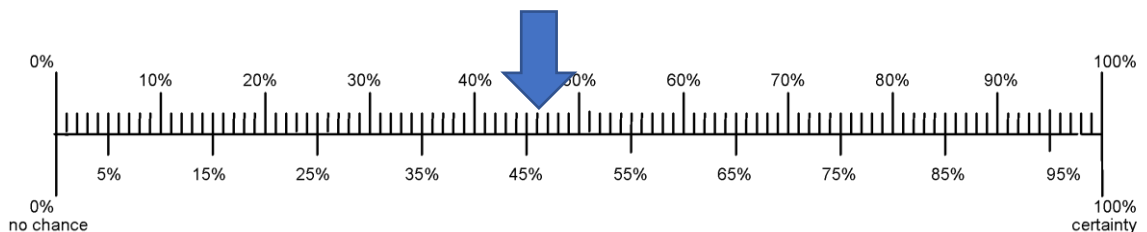
(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? \_\_\_\_\_ (minutes)

(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? \_\_\_\_\_(days)

#### Questionnaire 5

##### Instructions:

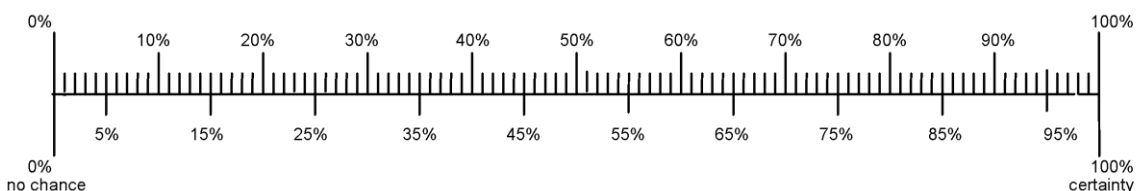
Each of these questions asks for your best guess at the chance that something will happen in the future. They use the “probability” scale that you see below. To answer each question, please put a mark on the scale at one specific tick mark, as follows:



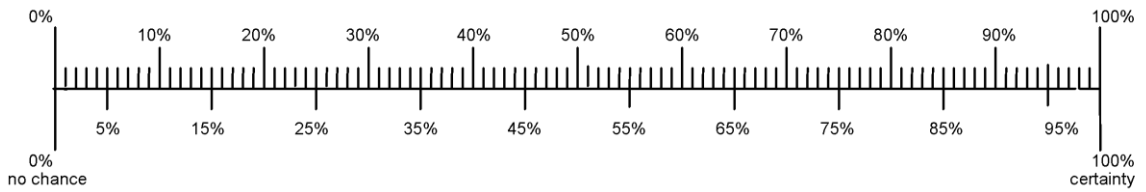
If you think that something has no chance of happening, mark it as having a 0% chance. If you think that something is certain to happen, mark it as having a 100% chance.

Just to make sure that you are comfortable with the scale, please answer the following practice questions.

**What is the probability that you will eat pizza during the next year?**



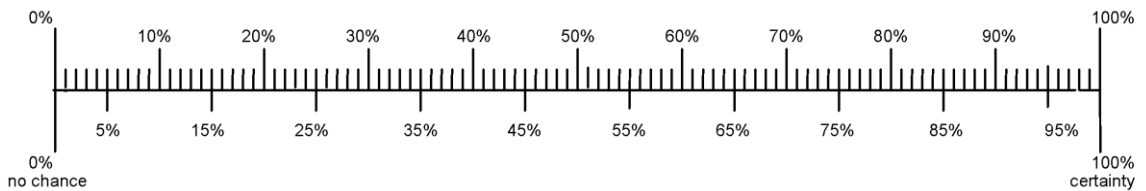
**What is the probability that you will get the flu during the next year?**



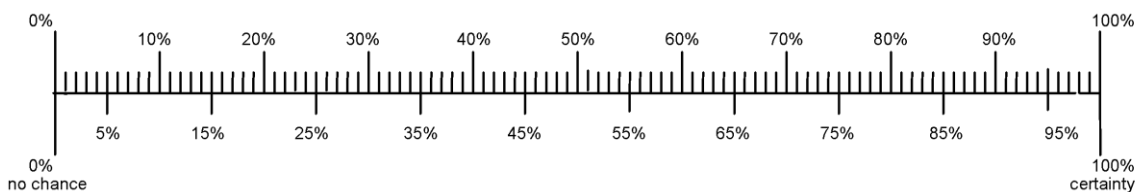
**That is the end of the practice. Please answer the following questions in the same way.**

**A. The following questions ask about events that may happen sometime during *the next year*.**

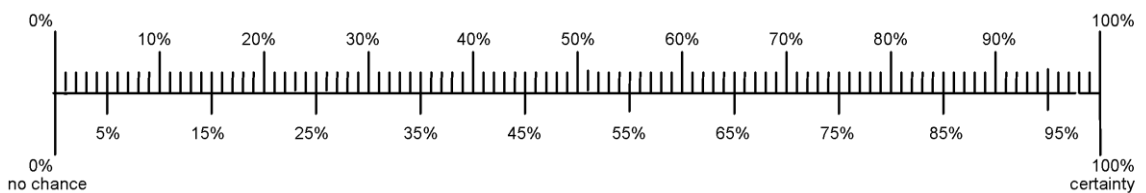
1. What is the probability that you will get into a car accident while driving during the next year?



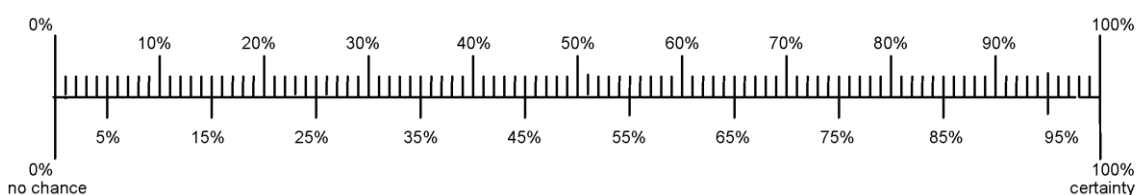
2. What is the probability that you will have a cavity filled during the next year?



3. What is the probability that you will die (from any cause -- crime, illness, accident, and so on) during the next year?

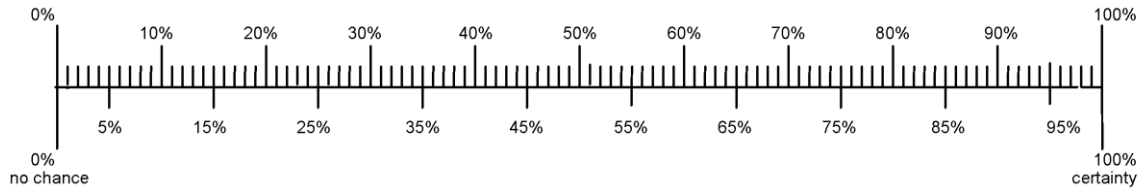


4. What is the probability that someone will steal something from you during the next year?

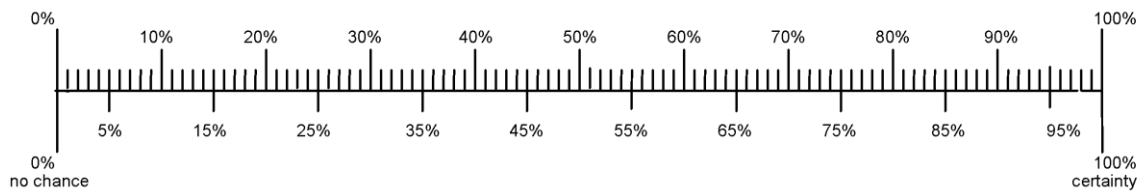




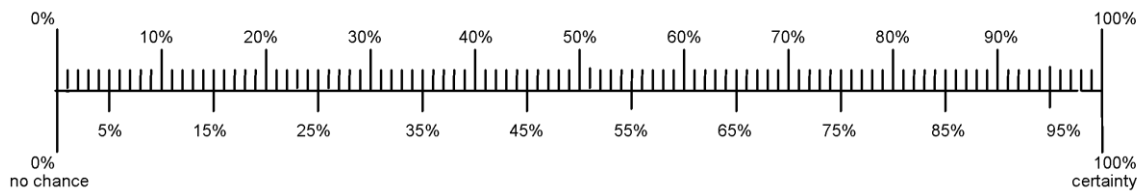
5. What is the probability that you will move your permanent address to another country some time during the next year?



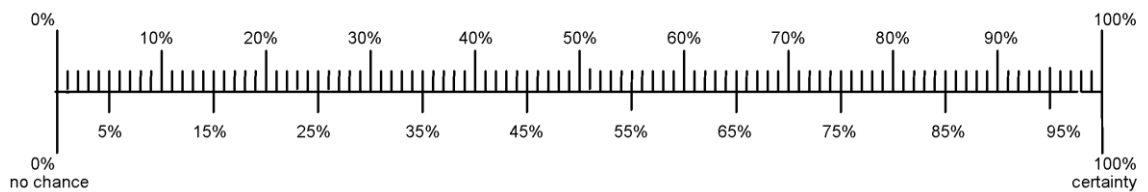
6. What is the probability that you will die in a terrorist attack during the next year?



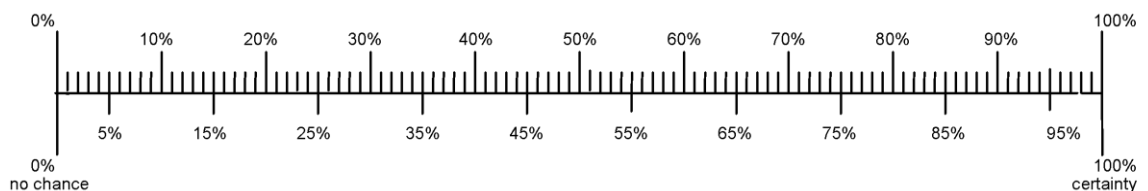
7. What is the probability that someone will break into your home and steal something from you during the next year?



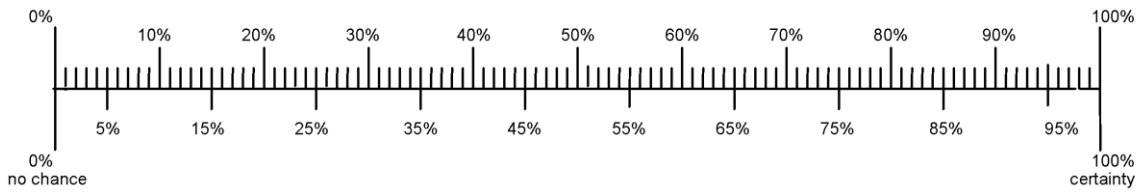
8. What is the probability that you will keep your permanent address in the same country during the next year?



9. What is the probability that you will visit a dentist, for any reason, during the next year?

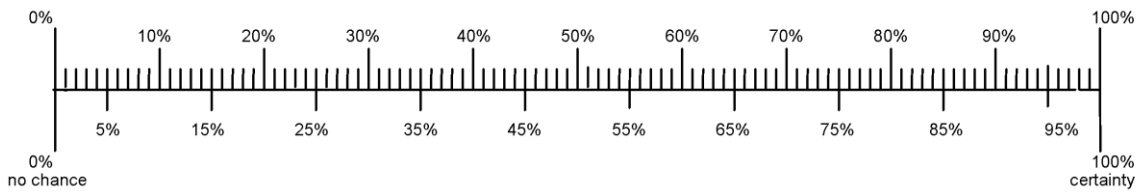


10. What is the probability that your driving will be accident-free during the next year?

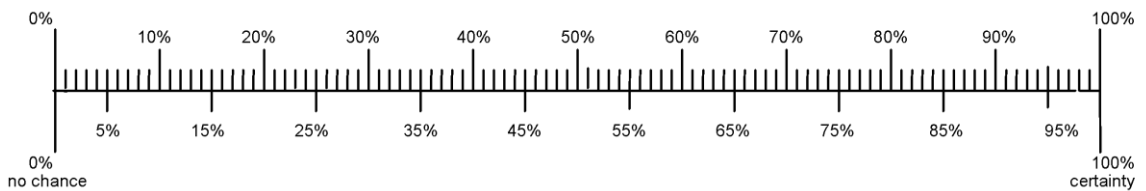


**B. The following questions ask about events that may happen sometime during *the next 5 years*.**

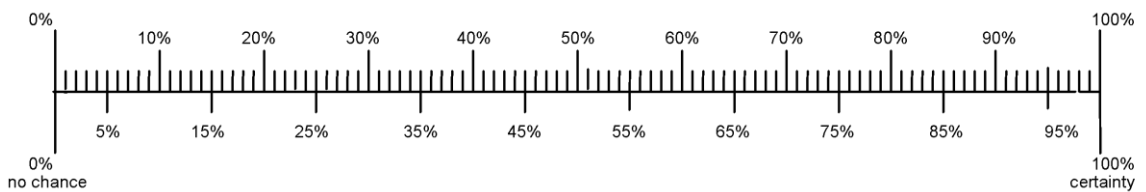
1. What is the probability that you will get into a car accident while driving during the next 5 years?



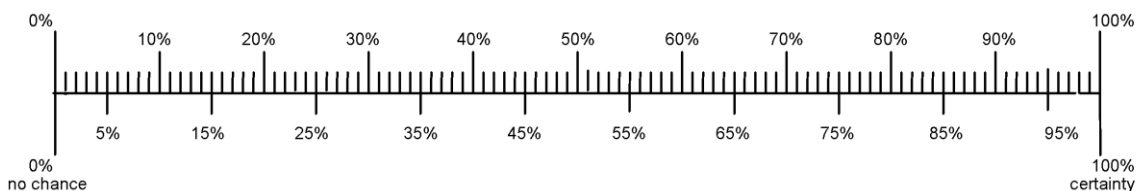
2. What is the probability that you will have a cavity filled during the next 5 years?



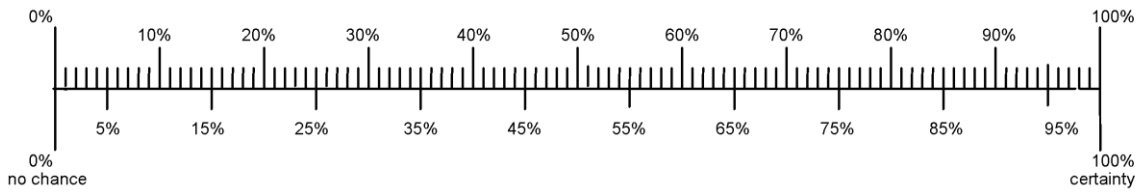
3. What is the probability that you will die (from any cause -- crime, illness, accident, and so on) during the next 5 years?



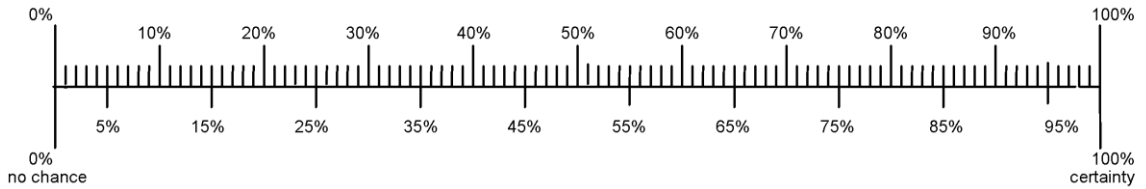
4. What is the probability that someone will steal something from you during the next 5 years?



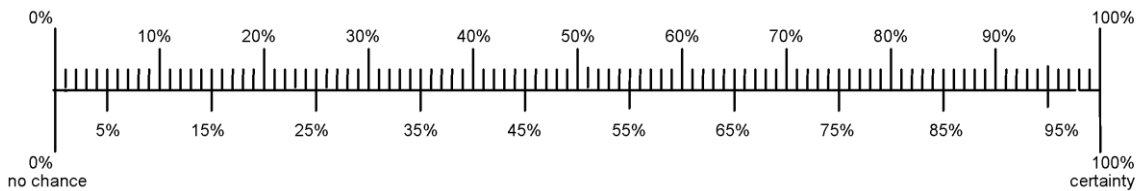
5. What is the probability that you will move your permanent address to another country some time during the next 5 years?



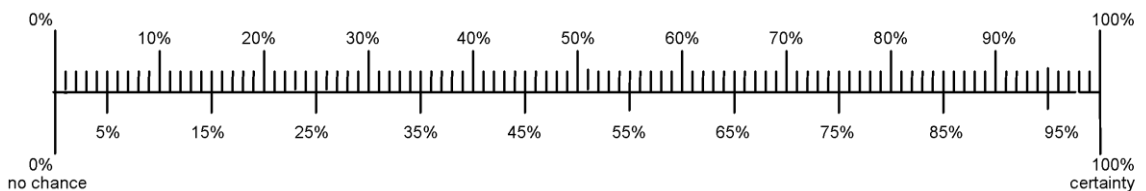
6. What is the probability that you will die in a terrorist attack during the next 5 years?



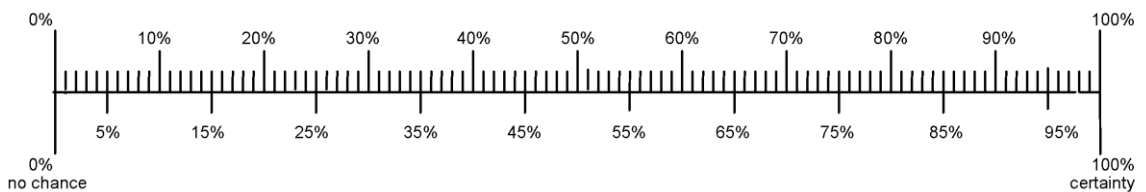
7. What is the probability that someone will break into your home and steal something from you during the next 5 years?



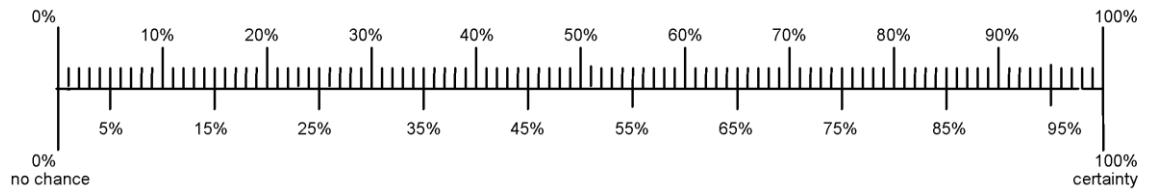
8. What is the probability that you will keep your permanent address in the same state during the next 5 years?



9. What is the probability that you will visit a dentist, for any reason, during the next 5 years?



10. What is the probability that your driving will be accident-free during the next 5 years?



### Questionnaire 6

#### Instructions :

The following pages contain phrases describing people's behaviours. Please use the rating scale next to each phrase to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then click the circle that corresponds to the accuracy of the statement.

	Very Inaccurate (1)	Moderately Inaccurate (2)	Neither Accurate Nor Inaccurate (3)	Moderately Accurate (4)	Very Accurate
Have a vivid imagination.					
Trust others.					
Believe in the importance of art.					
Like to tidy up.					
Often feel blue.					
Love to help others.					
Keep my promises.					
Love excitement.					
Love to read challenging material.					
Jump into things without thinking.					
Enjoy wild flights of fantasy.					
Believe that others have good intentions.					
See beauty in things that others might not notice.					
Often forget to put things back in their proper place.					
Dislike myself.					
Am concerned about others.					
Tell the truth.					
Seek adventure.					
Avoid philosophical discussions.					
Make rash decisions.					
Love to daydream.					
Trust what people say.					
Do not like poetry.					

Leave a mess in my room.					
Am often down in the dumps.					
Am not interested in theoretical discussions					
Break rules.					
Enjoy being reckless.					
Have difficulty understanding abstract ideas.					
Am not interested in other people's problems.					
Rush into things.					
Like to get lost in thought.					
Distrust people.					
Do not enjoy going to art museums.					
Leave my belongings around.					
Feel comfortable with myself.					
Take no time for others.					
Break my promises.					
Act wild and crazy.					
Act without thinking.					

### Questionnaire 7

	<b>1 = not at all true</b>	<b>2 = barely true</b>	<b>3 = somewhat true</b>	<b>4 = completely true</b>
<b>When solving my own problems other people's advice can be helpful.</b>				
<b>I try to talk and explain my stress in order to get feedback from my friends.</b>				
<b>Information I get from others has often helped me deal with my problems.</b>				
<b>I can usually identify people who can help me develop my own solutions to problems.</b>				
<b>I ask others what they would do in my situation.</b>				
<b>Talking to others can be really useful because it provides another perspective on the problem.</b>				
<b>Before getting messed up with a problem I'll call a friend to talk about it.</b>				

<b>When I am in trouble I can usually work out something with the help of others.</b>				
---	--	--	--	--

**Questionnaire 8**

	<b>1 = not at all true</b>	<b>2 = barely true</b>	<b>3 = somewhat true</b>	<b>4 = completely true</b>
<b>I imagine myself solving difficult problems.</b>				
<b>Rather than acting impulsively, I usually think of various ways to solve a problem.</b>				
<b>In my mind I go through many different scenarios in order to prepare myself for different outcomes.</b>				
<b>I tackle a problem by thinking about realistic alternatives.</b>				
<b>When I have a problem with my co-workers, friends, or family, I imagine beforehand how I will deal with them successfully.</b>				
<b>Before tackling a difficult task I imagine success scenarios.</b>				
<b>I take action only after thinking carefully about a problem.</b>				
<b>I imagine myself solving a difficult problem before I actually have to face it.</b>				
<b>I address a problem from various angles until I find the appropriate action.</b>				
<b>When there are serious misunderstandings with co-workers, family members or friends, I practice before how I will deal with them.</b>				
<b>I think about every possible outcome to a problem before tackling it.</b>				

**Questionnaire 9**

	1 (totally disagree)	2	3	4	5	6	7 (totally agree)
I try to be the best in the team.							
I put effort in to win							
I do my best to surpass any others							
I always attempt to do better than others							
I strive for first place							

### Questionnaire 10

	1 (totally disagree)	2	3	4	5	6	7 (totally agree)
I enjoy working in situations involving competition with others							
It is important to me to perform better than others on a task							
I feel that winning is important in both work and games							
I try harder when I am in competition with other people.							

### Questionnaire 11

	1 = not at all true	2 = barely true	3 = somewhat true	4 = completely true
I plan for future eventualities.				
Rather than spending every cent I make, I like to save for a rainy day.				
I prepare for adverse events.				
Before disaster strikes I am well-prepared for its consequences.				
I plan my strategies to change a situation before I act.				
I develop my job skills to protect myself against unemployment.				

I make sure my family is well taken care of to protect them from adversity in the future.				
I think ahead to avoid dangerous situations.				
I plan strategies for what I hope will be the best possible outcome.				
I try to manage my money well in order to avoid being destitute in old age.				

### Questionnaire 12

	1 = not at all true	2 = barely true	3 = somewhat true	4 = completely true
I often find ways to break down difficult problems into manageable components.				
I make a plan and follow it.				
I break down a problem into smaller parts and do one part at a time.				
I make lists and try to focus on the most important things first.				

### Questionnaire 13

	1 = Not at all true	2 = Hardly true	3 = Moderately true	4 = Exactly true
I can always manage to solve difficult problems if I try hard enough.				
If someone opposes me, I can find the means and ways to get what I want.				
It is easy for me to stick to my aims and accomplish my goals.				
I am confident that I could deal efficiently with unexpected events.				
Thanks to my resourcefulness, I know how to handle unforeseen situations.				
I can solve most problems if I invest the necessary effort.				
I can remain calm when facing difficulties because I can rely on my coping abilities.				



<b>When I am confronted with a problem, I can usually find several solutions.</b>				
<b>If I am in trouble, I can usually think of a solution.</b>				
<b>I can usually handle whatever comes my way.</b>				

#### Questionnaire 14

	<b>1 = not at all true</b>	<b>2 = barely true</b>	<b>3 = somewhat true</b>	<b>4 = completely true</b>
<b>I am a "take charge" person.</b>				
<b>I try to let things work out on their own.</b>				
<b>After attaining a goal, I look for another, more challenging one.</b>				
<b>I like challenges and beating the odds.</b>				
<b>I visualise my dreams and try to achieve them.</b>				
<b>Despite numerous setbacks, I usually succeed in getting what I want.</b>				
<b>I try to pinpoint what I need to succeed.</b>				
<b>I always try to find a way to work around obstacles; nothing really stops me.</b>				
<b>I often see myself failing so I don't get my hopes up too high.</b>				
<b>When I apply for a position, I imagine myself filling it.</b>				
<b>I turn obstacles into positive experiences.</b>				
<b>If someone tells me I can't do something, you can be sure I will do it.</b>				
<b>When I experience a problem, I take the initiative in resolving it.</b>				
<b>When I have a problem, I usually see myself in a no-win situation.</b>				

**Questionnaire 15**

	<b>(1) not at all true</b>	<b>(2) barely true</b>	<b>(3) moderately true</b>	<b>(4) exactly true</b>
<b>I can concentrate on one activity for a long time, if necessary.</b>				
<b>If I am distracted from an activity, I don't have any problem coming back to the topic quickly.</b>				
<b>If an activity arouses my feelings too much, I can calm myself down so that I can continue with the activity soon</b>				
<b>If an activity requires a problem-oriented attitude, I can control my feelings</b>				
<b>It is difficult for me to suppress thoughts that interfere with what I need to do</b>				
<b>I can control my thoughts from distracting me from the task at hand</b>				
<b>When I worry about something, I cannot concentrate on an activity</b>				
<b>After an interruption, I don't have any problem resuming my concentrated style of working</b>				
<b>I have a whole bunch of thoughts and feelings that interfere with my ability to work in a focused way</b>				
<b>I stay focused on my goal and don't allow anything to distract me from my plan of action</b>				

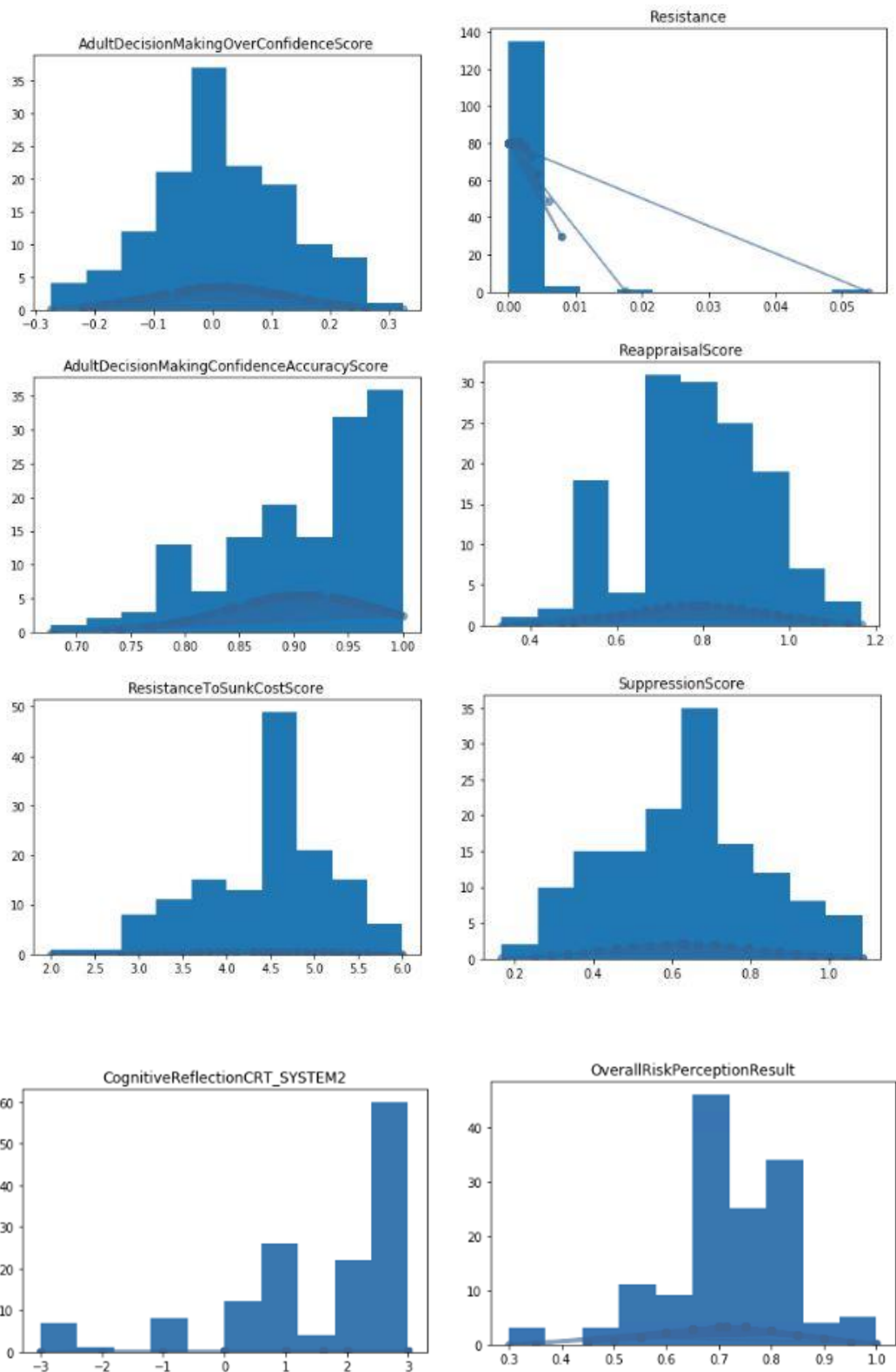
**Questionnaire 16**

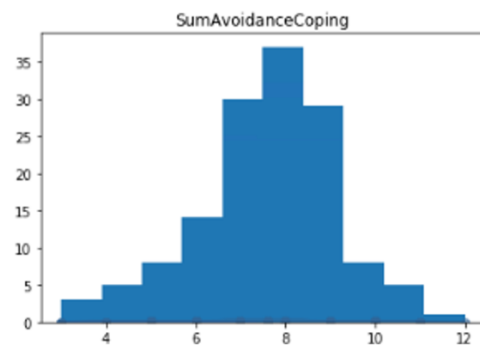
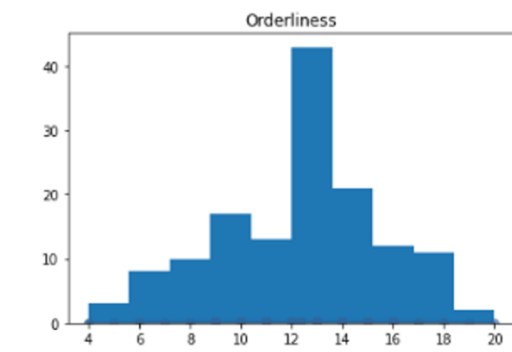
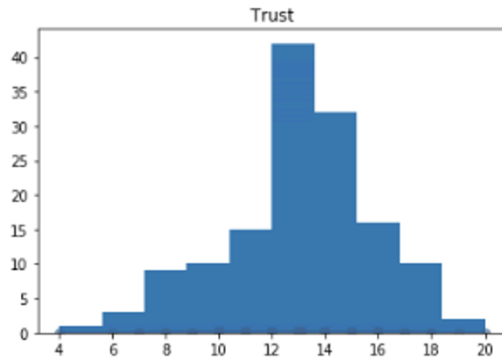
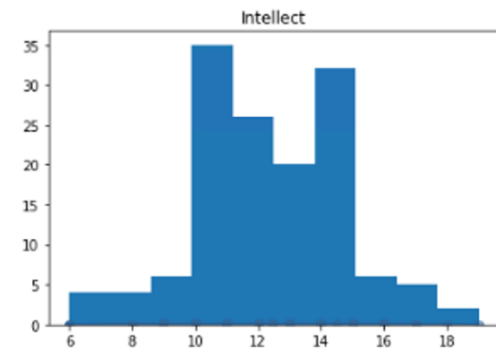
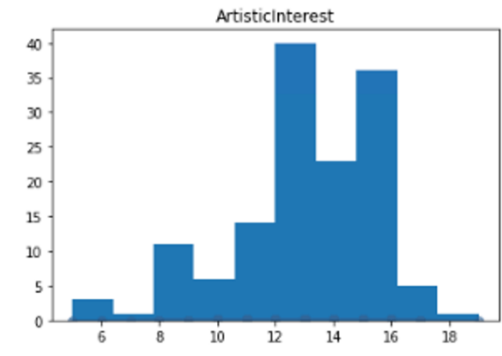
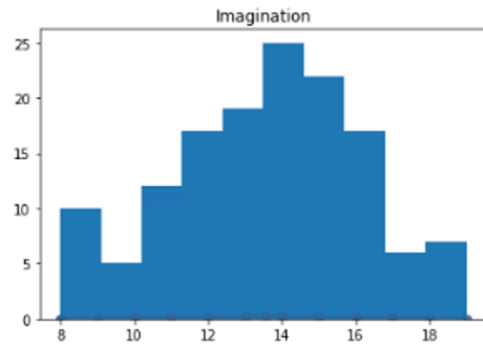
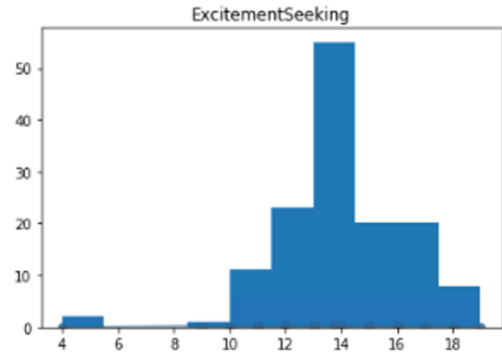
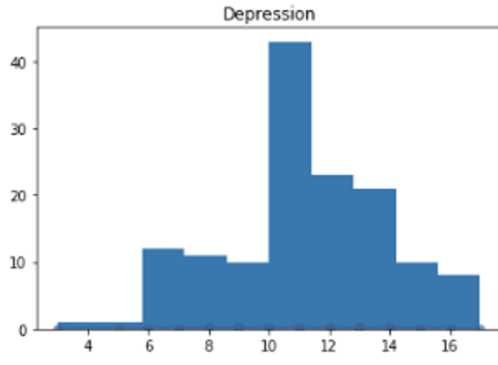
	<b>1 (totally disagree)</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7 (totally agree)</b>
<b>I hate competition.</b>							
<b>I find competition very tiresome.</b>							
<b>Competition makes me feel disgust.</b>							
<b>I think competition will destroy interpersonal harmony and cooperation.</b>							

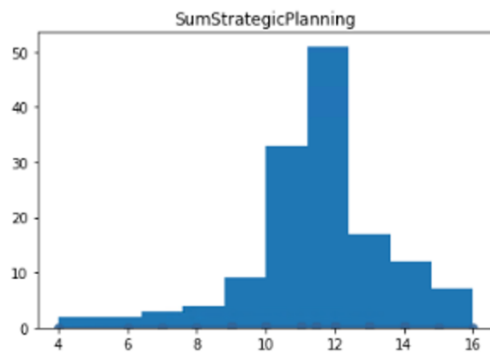
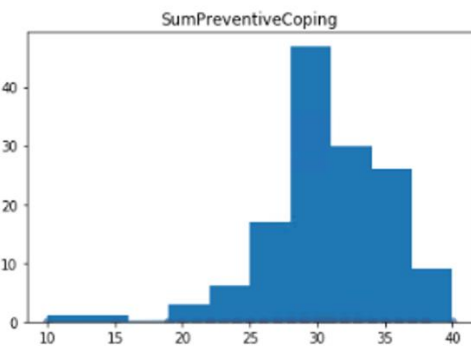
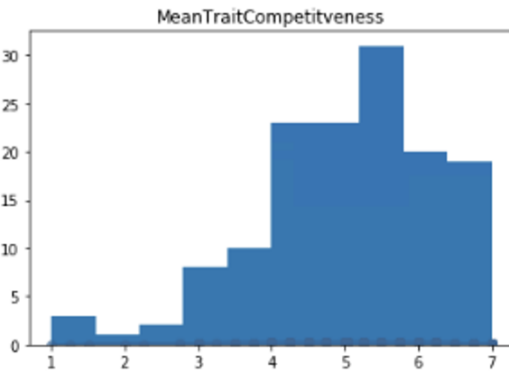
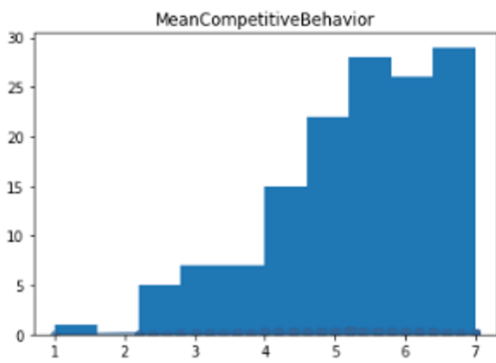
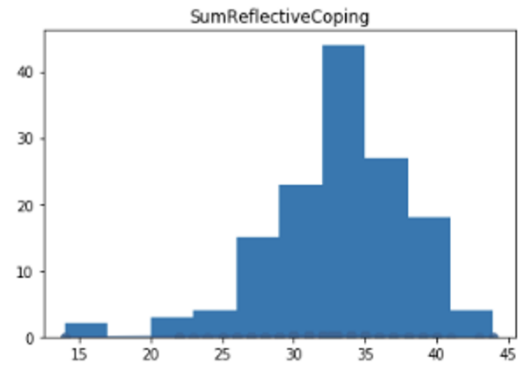
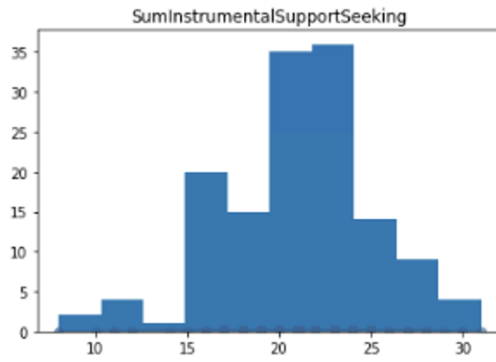
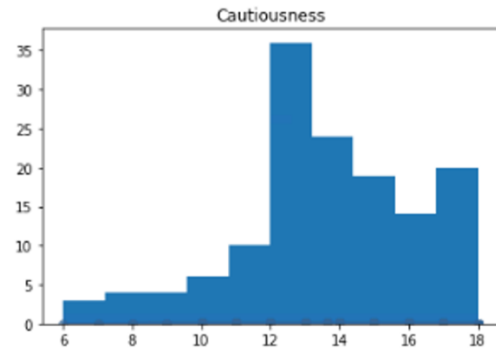
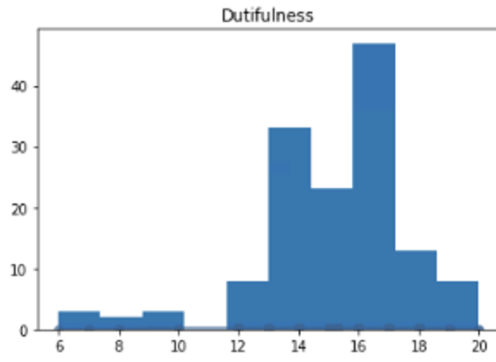
**Questionnaire 17**

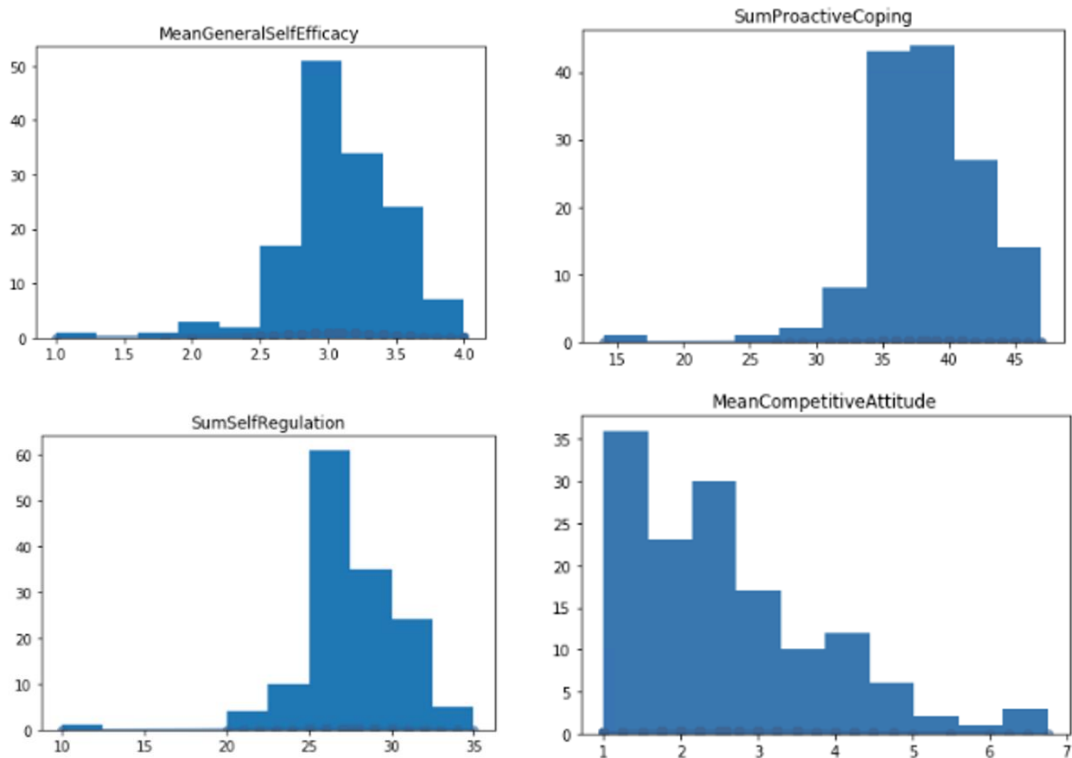
	<b>1 = not at all true</b>	<b>2 = barely true</b>	<b>3 = somewhat true</b>	<b>4 = completely true</b>
<b>When I have a problem I like to sleep on it.</b>				
<b>If I find a problem too difficult sometimes I put it aside until I'm ready to deal with it.</b>				
<b>When I have a problem I usually let it simmer on the back burner for a while.</b>				

### Appendix 3: Distribution of scores for each Feature used in Training Data









## Appendix 4: Informed consent form used in phase 2

### What we are doing?

As part of ----- and UCL's collaborative AI project, we are collecting data from ----- traders all around the world, to get better insights about behavioural patterns in trading and about traders' profiles.

### What we are asking you to do?

If you agree to take part, you will be asked to complete a questionnaire asking about your emotions, self evaluation and decision making. None of the tests evaluate your knowledge about trading or general intelligence.

### What happens to the information that you give us?

You do not have to take part in the survey if you do not want to. However, by completing it, you will help us to understand better the factors that are associated with trading performance and to improve ----- recruitment and training process. Any information that you give us will be treated in strict confidence, will be processed so long as it is required for this specific project, and will not be published or shared in any way that can be related to your name, or to any data identifying you. Only UCL's----- AI team will know that you have completed the questionnaire. The data that you will share with us in this research will be used for data analysis in the AI project. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the originally collected data.

### Do I have to take part?

Taking part in the study by completing this survey is entirely voluntary and the refusal to agree to participate will involve no penalty, nor will it affect your employment. If you change your mind about participating at any time over the course of the study, you can do so by contacting ---- or -----.

### **Consent Form**

Please read through the statements below and continue with the study if you agree with each statement.

1. I have had the opportunity to consider the information about this study.
2. I understand that my participation in this study is voluntary and that I can withdraw from the study at any time.
3. I understand that by agreeing to participate, I give consent for my survey responses to be matched with other data collected by ---- about My trading for the purpose of the UCL----- AI project. I understand that my information will be handled in accordance with all applicable data protection legislation.
4. I understand that my data will be confidential. It will not be possible to identify me in any publications.
5. I know how to contact the conducting team if I need to.

---

## **Appendix 5: Informed consent form used in phase 3**

### **[Transparency Framework for AI Products in Learning Contexts]**

#### **Participant Consent Form**

If you are happy to participate in this study please complete this consent form by ticking each item as appropriate, and return via email:

- 1) I confirm that I understand the context of this research and its uses, and have had my concerns adequately answered.
- 2) I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason.
- 3) I know that I can refuse to answer any or all of the questions and that I can withdraw from the interview at any point.



- 4) I agree for the interview to be video (and/or audio) recorded, and that recordings will be kept secure. I know that all data will be kept under the terms of the General Data Protection Regulation (GDPR).
- 5) I agree that small direct quotes may be used in reports.
- 6) I understand that in exceptional circumstances anonymity and confidentiality would have to be broken, for example, if it was felt that practice was putting children at risk, or there were concerns regarding professional misconduct. In these circumstances advice would be sought from a senior manager from another local authority who will advise on the appropriate course of action and as to whether we need to inform the authority of what you have told us.

Name:.....  
 ....

Signature: ..... Date:  
 .....

Name of researcher: Muhammad Ali Chaudhry

Signature:

## Appendix 6: Questions for Interviews in phase 3

### Stage 1 Questions:

- What is your background and job title?
- Are you using ed-tech products in your school?
- What are your views about AI in ed-tech?
- Did you and/or teachers received any training before the product was deployed?
  - If yes, were you told in which contexts this product does not work
- What do you do if there is a conflict between AI and your judgement?
  - Does this effect your confidence?
- Have you ever demanded or requested more transparency in AI tools in the past? Or do you mostly trust their judgement?
- (After sharing some details about the framework) Would you find a framework like this on AI Transparency useful?

- Would it enhance your understanding of AI products?
- Have you had such conversations on AI Ethics with ed-tech companies or with anyone else?
  - If yes, do you demand Transparency/Explanability from ed-tech companies?
- Would you find a framework like this on AI Transparency useful?
- Any recommendations on what you think can be improved in this framework?

### **Specific Questions for AI Practitioners**

- Do you take any measures to ensure adverse consequences of AI
- Have you thought about making the entire AI development pipeline transparent
- Do you get any specific requests for ethical AI or transparency in AI from educational institutions

### **Stage 2 Questions:**

#### **Educators:**

- Are you using in ed-tech products in your school?
  - Are they AI powered?
  -
- Did you and/or teachers received any training before the product was deployed?
  - In the training were you told when not to use or rely on this AI's recommendations?
- What do you do if there is a conflict between AI and your judgement?
  - Does this effect your confidence?
- Have you ever demanded or requested more transparency in AI tools in the past? Or do you mostly trust their judgement?
- Would you find a framework like this on AI Transparency useful?
  - Would it enhance your understanding of AI products?
- Do you think if this framework is added on teacher training, it would enhance teachers' understanding of the AI products they use?
- Have you had such conversations on AI Ethics before with anyone?
  - With whom
  - How often

#### **Ed-tech Experts:**

- What are your views about AI in ed-tech?
  - Is it impactful?
  - Is it over-hyped?
  - Does it reduce teacher workload?
  - Does it improve learning outcomes?
  
- Have you had such conversations on AI Ethics before with anyone?
  - With whom
  - How often
  
- Do you demand Transparency/Explanability from ed-tech companies?
  
- Do you think regulations like GDPR have any impact on AI Ethics? And Transparency?
  
- Do you know of any mishaps in AI in Education? Or AI going wrong?
  
- Would you find a framework like this on AI Transparency useful?
  - Would it enhance your understanding of AI products?
  - Do you think you can use this framework as an auditing tool?

## **AI Practitioners:**

- When you are developing AI products, do you get any Transparency requirements from clients? Like they want more insights into the data processing etc
  
- Do you usually make your models explainable?
  - If yes, how?
  - Which tools do you use?
  
- Would you find a framework like this on AI Transparency useful?
  - Would it enhance your understanding of AI products?
  - Do you think this framework would help with the documentation of your AI development?
  - Do you think this framework can be used as an auditing tool?
  
- Have you had such conversations on AI Ethics before with anyone?
  - With whom
  - How often
  
- Have you thought about AI transparency like this before, in terms of its application on every stage of the ML development pipeline?
  
- Have you had any conversations on AI Transparency with any one before?
  - With whom
  - How often
  
- Any recommendations on what you think can be improved in this framework?

## **Appendix 7: Ethics Approval from the UCL Research Ethics Committee**



## UCL Research Ethics Committee

**Note to Applicants:** It is important for you to include all relevant information about your research in this application form as your ethical approval will be based on this form. Therefore anything not included will not be part of any ethical approval.

*You are advised to read the Guidance for Applicants when completing this form.*

### Application For Ethical Review: Low Risk

Are you applying for an urgent accelerated review? Yes  No

If yes, please state your reasons below. Note: Accelerated reviews are for exceptional circumstances only and need to be justified in detail.

This is because I am submitting this new application because there is a slight change in the methodology of my research for which the ethics approval was gained earlier. I was recommended to apply for an urgent accelerated review rather than submit amendments to my current ethics approval, as processing of new application can be faster.

Is this application for a continuation of a research project that already has ethical approval? *For example, a preliminary/pilot study has been completed and is this an application for a follow-up project?* Yes  No

**If yes,** provide brief details (see guidelines) including the title and ethics id number for the previous study:

### Section A: Application details

1	<b>Title of Project</b>	Exploring transparency through a design framework in the AI powered ed-tech products
2	<b>Proposed data collection start date</b>	01 Sept 2020
3	<b>Proposed data collection end date</b>	31 July 2021
4	<b>Project Ethics Identification Number</b>	
5	<b>Principal Investigator</b> (*for student projects, your supervisor should be identified as the PI)	Professor Rose Luckin
6	<b>Position held</b>	Professor of Learner Centred Design
7	<b>Faculty/Department</b>	Culture, Communication and Media
9	<b>Contact Details</b> Email: Telephone:	
10	<b>Provide details of other Co-Investigators/Partners/Collaborators who will work on the project.</b> <i>Note: This includes those with access to the data such as transcribers.</i>	

Name: Muhammad Ali Chaudhry Position held: PhD Student Faculty/Department: Culture, Communication and Media Location (UCL/overseas/other UK institution): UCL Email: [REDACTED]	Name: Mutlu Cukurova Position held: Associate Professor Faculty/Department: Culture, Communication and Media Location (UCL/overseas/other UK institution): Email:
If you <b>do not know</b> the names of all collaborators, please write their roles in the research.	

<b>11 If the project is funded (this includes non-monetary awards such as laboratory facilities)</b>	
Name of Funder	Self-funded
Is the funding confirmed?	

<b>12 Name of Sponsor</b>
The Sponsor is the organisation taking responsibility for the project, which will usually be UCL. If the Sponsor is not UCL, please state the name of the sponsor.

<b>13 If this is a student project</b>	
Name	Muhammad Ali Chaudhry
Faculty/Department	Culture, Communication and Media
Position Held (please tick)	<input type="checkbox"/> Undergraduate/Bachelor project (if so, provide course title/number: _____) <input type="checkbox"/> Master's project (if so, provide course title/number: _____) <input checked="" type="checkbox"/> PhD <input type="checkbox"/> staff led research project which may involve one or more students
Contact details	Muhammad.Chaudhry.16@ucl.ac.uk

<b>Section B: Project details</b>
-----------------------------------

The following questions relate to the objectives, methods, methodology and location of the study. Please ensure that you answer each question in lay language.

<b>14 Provide a <i>brief</i> (300 words max) background to the project, including its intended aims.</b>
<p>This project aims to explore the issue of transparency through a design framework in the AI powered ed-tech products. The research question for this project is as follows: What kind of design framework can be applied to evaluate transparency of the AI powered ed-tech products that are used in learning contexts:</p> <ul style="list-style-type: none"> <li>• How can this framework be utilised to document any forms of correlations, causation or bias in the data used for developing AI products</li> <li>• How can this framework assist in filling the Awareness Gap (the gap between digital data and human experiences)</li> <li>• How can this framework be utilised by ed-tech companies to make their AI development process transparent for all the stakeholders involved</li> </ul>

The data used for developing this framework was collected by a financial services company, anonymised, and then shared for this research. After the framework is prepared, it is being tested with various educators, ed-tech experts and AI practitioners through interviews. This interviews data is being collected by the Phd candidate. This research would bridge the gap between Machine Learning (ML) and Learning Science communities in education. It would enhance Human-AI understanding through the ‘human in the loop’ by increasing their understanding of the data on which ML models are trained, potentially leading to more informed decisions. It will contribute towards ethical deployment of AI systems in real world.

**15 Methodology & Methods** (tick all that apply)

- |   |  |
|---|--|
| <input checked="" type="checkbox"/> Interviews*<br><input type="checkbox"/> Focus groups*<br><input type="checkbox"/> Questionnaires (including oral questions)*<br><input type="checkbox"/> Action Research<br><input type="checkbox"/> Observation<br>Participant Observation<br><input type="checkbox"/> Documentary analysis (including use of personal records)<br><input type="checkbox"/> Audio/visual recordings (including photographs)<br><i>*Attach copies to application (see below).</i> | <input type="checkbox"/> Collection/use of sensor or locational data<br><input type="checkbox"/> Controlled Trial<br><input type="checkbox"/> Intervention study (including changing environments)<br><input type="checkbox"/> Systematic review<br><input checked="" type="checkbox"/> Secondary data analysis – <b>(See Section D)</b><br><input type="checkbox"/> Advisory/consultation groups<br><input type="checkbox"/> Other, give details: |
|---|--|

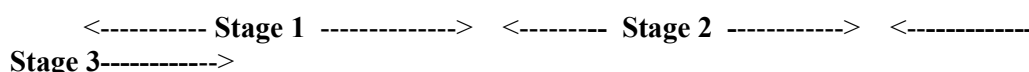
**16a Provide – in lay person’s language - an overview of the project;** focusing on your methodology and including information on what data/samples will be taken (including a description of the topics/questions to be asked), how data collection will occur and what (if relevant) participants will be asked to do. This should include a justification for the methods chosen. **(500 words max)**

Please **do not** attach or copy and paste a research proposal or case for support. This project aims to develop a framework to evaluate transparency of artificial intelligence implementations in ed-tech products. It will evaluate different aspects of transparency in all the stages of AI product development, including data processing stage, machine learning modelling stage and deployment of AI products in production. The main research questions this project aims to answer are as follows:

What kind of design framework can be applied to evaluate transparency of the AI powered ed-tech products that are used in learning contexts:

- How can this framework be utilised to document any forms of correlations, causation or bias in the data used for developing AI products
- How can this framework assist in filling the Awareness Gap (the gap between digital data and human experiences)
- How can this framework be utilised by ed-tech companies to make their AI development process transparent for all the stakeholders involved

Firstly, the framework is developed using secondary (anonymised) data from a financial services company. AI powered HR tool is developed for the financial services company using ‘human in the loop’ approach. While developing this tool, different aspects of transparency across various stages of development are analysed. Three different stages of AI tool development were identified, as shown in figure 1 below



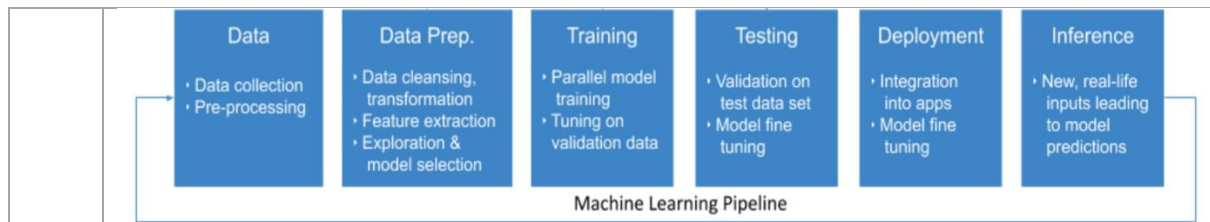


Fig 1: Machine Learning Pipeline<sup>18</sup>

Based on the three different stages of AI tool development in figure 1, a framework is developed that helps in evaluating how transparent are these stages and for whom are they transparent (general public or tech experts). After the framework is developed, it is tested with educators, ed-tech experts and AI practitioners working in ed-tech companies.

Testing of the framework is completed in two phases. In phase 1 interviews conducted with educators, ed-tech experts and AI practitioners, they are asked open-ended questions about AI hype, AI's role in education, ethics in AI and importance of transparency in AI powered ed-tech products. Some details about the framework are shared with them and their opinions are noted.

Framework is improved based on phase 1 interviews and then these improvements are verified from a different group of educators, ed-tech experts and AI practitioners in phase 2 interviews.

**16b**

**Attachments**

If applicable, please attach a copy of any interview questions/workshop topic guides/questionnaires/test (such as psychometric), etc and state whether they are in final or draft form.

Questions for phase 1 are as follows:

- What is your background and job title?
- Are you using ed-tech products in your school?
- What are your views about AI in ed-tech?
- Did you and/or teachers received any training before the product was deployed?
  - If yes, were you told in which contexts this product does not work
- What do you do if there is a conflict between AI and your judgement?
  - Does this effect your confidence?
- Have you ever demanded or requested more transparency in AI tools in the past? Or do you mostly trust their judgement?
- (After sharing some details about the framework) Would you find a framework like this on AI Transparency useful?
  - Would it enhance your understanding of AI products?
- Have you had such conversations on AI Ethics with ed-tech companies or with anyone else?
  - If yes, do you demand Transparency/Explanability from ed-tech companies?

<sup>18</sup> <https://www.agilestacks.com/tutorials/ml-pipelines>



- Would you find a framework like this on AI Transparency useful?
- Any recommendations on what you think can be improved in this framework?

### **Specific Questions for AI Practitioners**

- Do you take any measures to ensure adverse consequences of AI
- Have you thought about making the entire AI development pipeline transparent
- Do you get any specific requests for ethical AI or transparency in AI from educational institutions

Questions for phase 2 are as follows:

#### **Educators:**

- Are you using in ed-tech products in your school?
  - Are they AI powered?
  -
- Did you and/or teachers received any training before the product was deployed?
  - In the training were you told when not to use or rely on this AI's recommendations?
- What do you do if there is a conflict between AI and your judgement?
  - Does this effect your confidence?
- Have you ever demanded or requested more transparency in AI tools in the past? Or do you mostly trust their judgement?
- Would you find a framework like this on AI Transparency useful?
  - Would it enhance your understanding of AI products?
- Do you think if this framework is added on teacher training, it would enhance teachers' understanding of the AI products they use?
- Have you had such conversations on AI Ethics before with anyone?
  - With whom
  - How often

#### **Ed-tech Experts:**

- What are your views about AI in ed-tech?
  - Is it impactful?
  - Is it over-hyped?
  - Does is reduce teacher workload?
  - Does it improve learning outcomes?
- Have you had such conversations on AI Ethics before with anyone?
  - With whom
  - How often
- Do you demand Transparency/Explanability from ed-tech companies?

	<ul style="list-style-type: none"> <li>• Do you think regulations like GDPR have any impact on AI Ethics? And Transparency?</li> <li>• Do you know of any mishaps in AI in Education? Or AI going wrong?</li> <li>• Would you find a framework like this on AI Transparency useful? <ul style="list-style-type: none"> <li>○ Would it enhance your understanding of AI products?</li> <li>○ Do you think you can use this framework as an auditing tool?</li> </ul> </li> </ul> <p><b>AI Practitioners:</b></p> <ul style="list-style-type: none"> <li>• When you are developing AI products, do you get any Transparency requirements from clients? Like they want more insights into the data processing etc</li> <li>• Do you usually make your models explainable? <ul style="list-style-type: none"> <li>○ If yes, how?</li> <li>○ Which tools do you use?</li> </ul> </li> <li>• Would you find a framework like this on AI Transparency useful? <ul style="list-style-type: none"> <li>○ Would it enhance your understanding of AI products?</li> <li>○ Do you think this framework would help with the documentation of your AI development?</li> <li>○ Do you think this framework can be used as an auditing tool?</li> </ul> </li> <li>• Have you had such conversations on AI Ethics before with anyone? <ul style="list-style-type: none"> <li>○ With whom</li> <li>○ How often</li> </ul> </li> <li>• Have you thought about AI transparency like this before, in terms of its application on every stage of the ML development pipeline?</li> <li>• Have you had any conversations on AI Transparency with any one before? <ul style="list-style-type: none"> <li>○ With whom</li> <li>○ How often</li> </ul> </li> <li>• Any recommendations on what you think can be improved in this framework?</li> </ul>
--	--

<b>17</b>	<b>Please state which code of ethics (see Guidelines) will be adhered to for this research (for example, BERA, BPS, etc).</b>
	BERA

<b>Location of Research</b>	
<b>18</b>	<b>Please indicate where this research is taking place.</b> <input checked="" type="checkbox"/> UK only (Skip to 'location of fieldwork') <input type="checkbox"/> Overseas only <input type="checkbox"/> UK & overseas
<b>19</b>	<b>If the research includes work outside the UK, is ethical approval in the host country (local ethical approval) required? (See Guidelines.)</b>

	<p>Yes <input type="checkbox"/> No <input type="checkbox"/></p> <p><b>If no</b>, please explain why local ethical approval is not necessary.  <b>If yes</b>, provide details below including whether the ethical approval has been received.  <b>Note:</b> Full UCL ethical approval will not be granted until local ethical approval (if required) has been evidenced.</p>
<b>20</b>	<p><b>If you (or any members of your research team) are travelling overseas in person are there any concerns based on governmental travel advice (<a href="http://www.fco.gov.uk">www.fco.gov.uk</a>) for the region of travel?</b> Yes <input type="checkbox"/> No <input type="checkbox"/></p> <p><b>Note:</b> Check <a href="http://www.fco.gov.uk">www.fco.gov.uk</a> and submit a travel insurance form to UCL Finance (see application guidelines for more details). This can be accessed here: <a href="https://www.ucl.ac.uk/finance/secure/fin_acc/insurance.htm">https://www.ucl.ac.uk/finance/secure/fin_acc/insurance.htm</a> (You will need your UCL login details.)</p>

<b>21</b>	<p><b>State the location(s) where the research will be conducted and data collected. For example public spaces, schools, private company, using online methods, postal mail or telephone communications.</b></p> <p>Due to Covid'19, online meeting tools like Zoom and Microsoft Teams are being used to conduct the interviews for this research.</p>
<b>22</b>	<p><b>Does the research location require any additional permissions (e.g. obtaining access to schools, hospitals, private property, non-disclosure agreements, access to biodiversity permits (CBD), etc.)?</b></p> <p>Yes <input type="checkbox"/> No <input type="checkbox"/></p> <p><b>If yes</b>, please state the permissions required.</p>
<b>23</b>	<p><b>Have the above approvals been obtained?</b> Yes <input type="checkbox"/> No <input type="checkbox"/></p> <p><b>If yes</b>, please attach a copy of the approval correspondence.</p> <p><b>If not</b>, confirm they will be obtained prior to data collection. Yes <input type="checkbox"/> No <input type="checkbox"/></p>

### Section C: Details of Participants

In this form 'participants' means human participants and their data (including sensor/locational data, observational notes/images, tissue and blood samples, as well as DNA).

<b>24</b>	<b>Does the project involve the recruitment of participants?</b>
<b>Yes</b> <input checked="" type="checkbox"/>	Complete all parts of this Section.
<b>No</b> <input type="checkbox"/>	Move to Section D.

#### Participant Details

<b>25</b>	<p>Approximate maximum number of participants required: 20</p> <p>Approximate upper age limit: 55 Lower age limit: 22</p> <p>Justification for the age range and sample size: I am interviewing educators, ed-tech experts and AI practitioners who are mostly within this age group.</p>
-----------	---

#### Recruitment/Sampling

<b>26</b>	Describe how potential participants will be recruited into the study.
-----------	---

	<p><b>Note:</b> This should include reference to how you will identify and approach participants. For example, will participants self-identify themselves by responding to an advert for the study or will you approach them directly (such as in person or via email)?</p> <p>I will approach them directly myself through digital media (emails, LinkedIn etc). I am also relying on referrals from the participants I recruit initially.</p>
<b>Informed Consent</b>	
<b>27a</b>	<p>Describe the process you will use when seeking to obtain consent.</p> <p><b>Note:</b> This should include reference to what participants are being asked to consent to, such as whether their contribution will be identifiable/anonymous, limits to confidentiality and whether their data can be withdrawn at a later date.</p> <p><i>(An annotated template information sheet and consent form have been provided for your use.)</i></p> <p>Firstly, consent is sought to participate in this research by providing a summary of the research in the recruitment emails. After the participants have provided consent to take part in the interviews, consent forms to use participants views in the research are shared. For some participants, consent forms were shared after the interviews</p>
<b>27b</b>	<p><b>Attachments</b> Please list them below:</p> <p><i>Ensure that a copy of all recruitment documentation (recruitment emails/posters, information sheet/s, consent form/s) have been attached to the application.</i></p>
<b>27c</b>	<p>If you are <b>not</b> intending to seek consent from participants, clarify why below:</p>

<b>28</b>	<p>How will the results be disseminated (including communication of results with participants)?</p> <p>Final thesis of this research will be shared with participants after publishing.</p>
-----------	---

#### Section D: Accessing/Using Pre-collected Data

<b>Access to data</b>	
<b>29</b>	<p>If you are using data or information held by third party, please explain how you will obtain this. You should confirm that the information has been obtained in accordance with the General Data Protection Regulation 2018.</p> <p>Initially, the framework is developed based on the financial services company's data. This is secondary data collected by the company and anonymised before giving us access. I will access this data by logging in the remote PC of the company who has collected this data from its employees (OSTC). This data is stored on their servers. It was collected in accordance with the guidelines of GDPR.</p> <p>To test the effectiveness of the proposed framework, I will conduct interviews of educators, ed-tech experts and AI practitioners. The questions I will ask in these interviews are given above.</p>

<b>Accessing pre-collected data</b>	
<b>30</b>	<p><b>Does your study involve the use of previously collected data?</b></p> <p>No <input type="checkbox"/> Move to Section E.</p> <p>Yes <input checked="" type="checkbox"/> Complete all parts of this Section. <b>Note:</b> If you ticked any boxes with an asterisk (*), ensure further details are provided in Section E: Ethical Issues.</p>

<b>31</b>	<b>Name of dataset/s:</b>
-----------	---------------------------

<b>32</b>	<b>Owner of dataset/s (if applicable):</b> OSTC	
<b>33</b>	<b>Is the data in the public domain?</b> Yes <input type="checkbox"/> No <input checked="" type="checkbox"/> <b>If not, do you have the owner's permission/license?</b> Yes <input checked="" type="checkbox"/> No* <input type="checkbox"/>	
<b>33</b>	<b>Is the data anonymised?</b> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> <b>If not:</b> i. Do you plan to anonymise the data? Yes <input type="checkbox"/> No* <input type="checkbox"/> ii. Do you plan to use individual level data? Yes* <input type="checkbox"/> No <input type="checkbox"/> iii. Will you be linking data to individuals? Yes* <input type="checkbox"/> No <input type="checkbox"/>	
<b>34</b>	Is the data sensitive?	Yes* <input type="checkbox"/> No <input checked="" type="checkbox"/>
<b>35</b>	Will you be conducting analysis within the remit it was originally collected for?	Yes <input checked="" type="checkbox"/> No* <input type="checkbox"/>
<b>36</b>	If not, was consent gained from participants for subsequent/future analysis?	Yes <input type="checkbox"/> No* <input type="checkbox"/>

### Section E: Ethical Issues

<b>Ethical Issues</b>	
<b>37</b>	<p>Please address clearly any ethical issues that may arise in the course of this research and how they will be addressed. Further information and advice can be found in the guidelines. <b>Note:</b> All ethical issues should be addressed - <b>do not leave this section blank</b>. All projects give rise to ethical issues. If you think there are no ethical issues, you need to provide an explanation as to why.</p> <p>One of the ethical issues that can potentially arise in this research can be due to my bias to prove the effectiveness of the framework. To counteract this issue firstly I will seek feedback from independent researchers working on this project. Secondly, I will conduct a number of interviews to evaluate the impact of this framework. I will report the outcome of these studies with the research as well.</p> <p>Another ethical issue can be the embarrassment for some interviewees because they might think that they do not know enough about Artificial Intelligence and Machine Learning. To avoid this, I have kept the interview questions semi-structured and will make sure to provide extra details and explanations to interviewees who do not have a tech background.</p>

### Risks & Benefits

<b>38</b>	Please state any <i>benefits</i> to participants in taking part in the study (this includes feedback, access to services or incentives),
-----------	--

	<p>This research will directly or indirectly potentially benefit all the stakeholders participating in this research</p> <p>Benefits for Educators:</p> <ul style="list-style-type: none"> <li>• Safety from AI systems</li> <li>• Better understanding of what they need from an AI system</li> <li>• Awareness of what they currently have from AI systems</li> <li>• More informed decision making based on AI's predictions/recommendations</li> </ul> <p>Benefits for AI Practitioners and ed-tech companies:</p> <ul style="list-style-type: none"> <li>• Auditing of their AI implementation</li> <li>• Evaluation of their product from Ethical AI perspective</li> <li>• Recommendations for making their AI product more transparent</li> <li>• Better understanding of the assumptions made in the AI development</li> </ul> <p>This research would indirectly also benefit the candidates whose secondary data I will be using. The mentoring tool that will be developed using this data would optimise traders' performance and provide them instant feedback.</p>
39	<p>Do you intend to offer incentives or compensation, including access to free services)?</p> <p>Yes <input type="checkbox"/> No <input checked="" type="checkbox"/></p> <p><b>If yes, specify the amount to be paid and/or service to be offered as well as a justification for this.</b></p>
40	<p>Please state any <i>risks</i> to participants and how these risks will be managed.</p> <p>There are no risks for participants whose data is used in this research.</p>
41	<p>Please state any <i>risks</i> to you or your research team and how these risks will be managed.</p> <p>There are no risks for the research team who are collecting data in this research.</p>

## Section F: Appropriate Safeguards, Data Storage & Security

Please ensure that you answer each question and include all hard and electronic data.

42	<p><b>Will the research involve the collection and/or use of personal data?</b></p> <p>Yes <input checked="" type="checkbox"/> No <input type="checkbox"/></p> <p><i><b>Personal data</b> is data which relates to a living individual who can be identified from that data OR from the data and other information that is either currently held, or will be held by the data controller (the researcher).</i></p> <p><i>This includes:</i></p> <ul style="list-style-type: none"> <li>– any expression of opinion about the individual and any intentions of the data controller or any other person toward the individual.</li> <li>– sensor, location or visual data which may reveal information that enables the identification of a face, address, etc (some postcodes cover only one property).</li> <li>– combinations of data which may reveal identifiable data, such as names, email/postal addresses, date of birth, ethnicity, descriptions of health diagnosis or conditions, computer IP address (if relating to a device with a single user).</li> </ul>
----	--

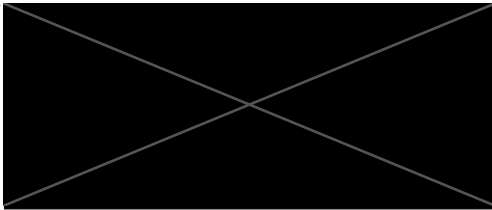
	<p>The only personal data I am collecting is participants' names and their designations / job titles. I am also collecting their opinions on the framework developed in this research, ed-tech use in schools and potential for AI in education</p> <p>If you do not have a registration number from Legal Services, please clarify why not: I have applied for registration with the Data Protection Office and am awaiting the Data Protection Registration number.</p>
<b>43</b>	<p><b>Is the research collecting or using</b></p> <ul style="list-style-type: none"> <li>– special category data as defined by the General Data Protection Regulation and/or</li> <li>– data which might be considered sensitive in some countries, cultures or contexts.</li> </ul> <p><b>If yes</b>, state whether explicit consent will be sought for its use and what data management measures are in place to adequately manage and protect the data.</p> <p>No.</p>

<b>44</b>	<p><b>All research projects using personal data must be registered with Legal Services before the data is collected, please provide the Data Protection Registration Number: Z6364106/2021/06/258</b></p> <p>If you do not have a registration number from Legal Services, please clarify why not:</p>
-----------	--

<b>During the project (including the write up and dissemination period)</b>	
<b>45</b>	<p><b>State what types of data will be generated from this project</b> (i.e. transcripts, videos, photos, audio tapes, field notes, etc).</p> <p>Videos and Notes.</p> <p><b>How will data be stored, including where and for how long?</b> This includes all hard copy and electronic data on laptops, share drives, usb/mobile devices.</p> <p>The data will be stored in Microsoft OneDrive as it is encrypted, when/if the data is downloaded to a laptop/hard drive, it will be kept in encrypted folders. The data will be stored for around 5 years for future analysis, references, corrections, requests from data providers etc.</p> <p><b>Who will have access to the data, including advisory groups and during transcription?</b></p> <p>Only the PhD student who collected this data will have access to it.</p>
<b>46</b>	<p><b>Do you confirm that all personal data will be stored and processed in compliance with the General Data Protection Regulation (GDPR 2018).</b></p> <p>Yes <input checked="" type="checkbox"/> No <input type="checkbox"/></p> <p><b>If not</b>, please clarify why.</p>
<b>47</b>	<p><b>Will personal data be processed or be sent outside of the European Economic Area (EEA)?*</b></p> <p>Yes <input type="checkbox"/> No <input checked="" type="checkbox"/></p> <p><b>If yes</b>, please confirm that there are adequate levels of protection in compliance with the GDPR 2018 and state what the arrangements are below.</p>

<b>After the project</b>	
<b>48</b>	<b>What data will be stored and how will you keep it secure?</b>

	Notes from the interviews conducted with participants. <b>Where will the data be stored and who will have access?</b> It will be stored on Microsoft OneDrive as it is encrypted.  <b>Will the data be securely deleted?</b> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> <b>If yes, please state when this will occur:</b> December 2027
49	<b>Will the data be archived for use by other researchers?</b> Yes <input type="checkbox"/> No <input checked="" type="checkbox"/> <b>If yes, please provide further details including whether researchers outside the European Economic Area will be given access.</b>

<b>Section G: Declaration to be Signed by the Principal Researcher</b> <b>I confirm that the information in this form is accurate to the best of my knowledge.</b>	
<b><u>For staff project:</u></b> Signature	
Date	
<b><u>For student project:</u></b> <b>I have met with and advised the student on the ethical aspects of this project design.</b>	
Signature	
Date:	30 June 2021

<b>Signature of your Head of Department (or Chair of your Departmental Ethics Committee or Departmental Ethics Lead)</b>
<b>Part A</b> I have read the 'criteria of minimal risk' as defined on page 3 of the Guidelines ( <a href="http://ethics.grad.ucl.ac.uk/forms/guidelines.pdf">http://ethics.grad.ucl.ac.uk/forms/guidelines.pdf</a> ) and I recommend that this application be considered by the Chair of the UCL REC. Yes <input type="checkbox"/> No <input type="checkbox"/>
<b>Part B</b> <b>I have discussed this project with the principal researcher who is suitably qualified to carry out this research and I approve it. I am satisfied that** (highlight as appropriate):</b>  <ol style="list-style-type: none"> <li><b>Data Protection registration:</b> <ul style="list-style-type: none"> <li>has been satisfactorily completed</li> <li>has been initiated</li> <li>is not required</li> </ul> </li> <li><b>A risk assessment:</b> <ul style="list-style-type: none"> <li>has been satisfactorily completed</li> <li>has been initiated</li> </ul> </li> <li><b>Appropriate insurance arrangements are in place and appropriate sponsorship [funding] has been approved and is in place to complete the study.</b></li> </ol> Yes <input type="checkbox"/> No <input type="checkbox"/>




**4. A Disclosure and Barring Service check(s):**

- has been satisfactorily completed
- has been initiated
- **is not required**

**Note:** Links to details of UCL's policies on the above can be found at:

<http://ethics.grad.ucl.ac.uk/procedures.php>

**\*\*If any of the above checks are not required please clarify why below.**

Name:	Muhammad Ali Chaudhry
Signature:	
Date:	18 <sup>th</sup> June 2021

Updated March 2019