Big Data and in online education: Who produces value and who reaps the rewards?

Student ID Number:

Word Count: 21,812

University of Glasgow, University of Malta, Tallinn University, Open University of Cyprus

August 16th, 2021

Abstract

From classifying learners to predicting learner behavior, the application of Big Data in online education has been vast. Besides the potential benefits of Big Data in education, it is necessary to critically engage with some ethical and social challenges that Big Data presents to the field of e-learning. The increasing use of big data by large institutional actors and corporations raises questions not only about data privacy and ownership, but whether this data is used to genuinely improve learner and teacher e-learning experiences, or solely for commercial profits and institutional benefits. When addressing ethical concerns regarding the use of Big Data in education, critiques often follow a reasoning that is in line with corporate interests and neoliberal logic of marketization of education. Given the importance of the pursuit for democratic online education, the need for critical perspectives in the field is ever-more essential. This research tries to critically address the role and impact of Big Data on labour relations and economic fairness in online education by examining both corporate and institutional data practices in e-learning. The presentation will put forward a provisional theory of the use of Big Data in two large e-learning platforms (Coursera and Blackboard) using critical grounded theory. The core category of *Exploitation of the learning community*, the three constituent concepts; *the Vendor-Institutional Complex, Use of learner generated value for profit,* and *the Behavioral monitoring and engineering*; and the sustaining category, *the Magic Trick*, were the foundational blocks for developing an emancipatory that addressed ethical issues of economic fairness regarding the use if big data in online education.

Acknowledgments

I would like to express my gratitude and appreciation for the following people. You were a vital source of love, comfort, motivation and inspiration for me during this process.

I thank my parents, Danijela and Gjorgji, and my sister Sara. Фала за тоа што секој ден ме обвивате со љубов и поддршка. Ве сакам!

I thank Ketevan. You are the most loving, caring, and inspiring companion I could ask for. I apricate you, your loyalty, and your love.

I thank Bonnie, my peers, and all the people involved in the IMAESC programme. I appreciate all the inspirational moments, interesting conversations, and caring experiences with you all.

I thank my supervisors Maria and Triin for their guidance and support. Without you and your support this would have been a much more difficult and tiring process.

Table of Contents

List of Tables

List of Figures

List of Acronyms

GTM – *Grounded Theory Methodology*

GT – *Grounded Theory*

CGT – *Critical Grounded Theory*

B2B – *Business-to-business*

AWS – *Amazon Web Services*

LA – *Learning Analytics*

CEO – *Chief Executive Officer*

Chapter 1: Introduction

The rapid technological advancement in computing in the past three decades has allowed humans to quickly and efficiently gather, access, and process large quantities of information. This revolution or breakthrough in information technology is often referred to as the Big Data Revolution (Kitchin, 2014). From cancer diagnosis and gene sequencing in biomedicine to predicting earthquakes in seismology, the contributions of big data to the sciences are extraordinarily valuable (Groves et al., 2016; Reyes, et al., 2013; Libbrecht & Noble, 2015). However, the use of big data has also been met with ethical concern and ambiguity, especially when human behaviour is the source of data, and political or economic benefit is the goal of data processing. Such cases include the case of Cambridge Analytica and the 2016 U.S. Presidential election, or the multiple privacy concerns over behavioural advertising conducted by big tech companies such as Facebook, Amazon, and Google (Bodle, 2016, González, 2017).

Just like many other industries, sciences, and areas of social life, education too, is under a mass wave of digitization and datafication. Meaning, more and more learning and teaching is done online, using software programs that run on, collect, and process massive amounts of digital data. Thus, the practices and logic of the big data revolution also penetrated education. The applications and uses of big data in online adult education are vast and various. Examples include predicting dropout rates, improving test scores, increasing online course engagement, or simply assistance with administrative tasks (Daniel, 2017; Huda et al., 2017; Liang et al., 2016; O'Reilly & Veeramachaneni, 2014; Prinsloo et al., 2015). Both in the industry and the academic literature, big data is mostly praised for its benefits and potential for improving the online education experience. However, concerns regarding privacy, data ownership, and individuality have also been on the rise (Chen & Liu, 2015; Johnson, 2014; Prinsloo & Slade, 2017; Williamson, 2017b).

To truly address the ethical concerns regarding big data in education, it is important to recognise the unprecedented nature of the big data revolution. It is unprecedented because it provides the opportunity to concentrate wealth, knowledge, and power like never before. Therefore, it is crucial to

approach the issue with openness and as few preconceptions as possible. Thus, it is important to not only raise questions regarding the commonly mentioned ethical concerns in online education, such as privacy and data security but also to adopt an approach that is open to examining and studying other, previously overlooked or neglected issues.

This work deals with exactly these rarely addressed ethical concerns. To conduct the study, I carried out a qualitative critical grounded theory case study of two of the most prominent online education providers, Coursera and Blackboard. The study resulted in a provisional conceptualization of the economic model of online education in the age of big data and the ethical concerns relating to it.

Statement of the Research Problem

Due to the fact that this thesis follows a grounded theory methodology (GTM), the research problem was not formulated before the data collection and analysis. In fact, according to Glaser, in a GTM study, the research problem is not a pre-set statement that identifies "the phenomenon to be studied" prior to data collection, but it is a product that emerges alongside the data collection and analysis processes (1992, p.22). Therefore, for this thesis, the research problem can be seen as a combination of two components. First, the selected area of concern or field of study before data collection and analysis, and second, the problems and gaps that emerged during data collection and analysis, guiding the process of theoretical sampling, which is discussed in more detail in Chapter 3.

Prior to data collection, the selected area of concern was developed through my previous observations and research experience, my critical positioning as a researcher, my previous engagement with the existing literature on big data in online education, and past work on algorithmic fairness and digital capitalism. Firstly, by way of observation and research experience in the field of online adult education, I have come to realize the crucial role of big data in online education. Furthermore, I have come to realise the unprecedented logic and mode of practice that came to dominate the socioeconomic relationships on the internet after the big data revolution. I will speak more on this unprecedented nature later on in the second chapter when discussing big data as a key concept for the thesis.

Secondly, my ontological and philosophical stance as a critical realist has also contributed to the development of the problem statement. Specifically, as a critical realist researcher, my field of study is concerned with uncovering 'generative mechanisms' that are not observable and hidden from perception (Belfrage and Hauf, 2017). Generative mechanisms are entities that explain why observable events or phenomena occur (Blom and Moren, 2011). Therefore, seeking to uncover veiled mechanisms, this thesis is emancipatory in nature and seeks to invoke change. It follows a critical approach, examining the less addressed and often overlooked aspects of big data in online education.

Lastly, the domain of study for this thesis is shaped by previous studies and literature. This literature can be split into two groups: literature relating to big data and the socio-economic relationships on the internet more broadly, and literature relating to the use of big data in online education. The former group covers multiple works dealing with ethical concerns regarding big data, digital labour, the commodification of information, and the concentration of knowledge on the internet. Most notably, these works include Zuboff (2019), Fuchs (2012), Srnicek (2017a). The latter group covers research and work that deals with ethical concerns regarding the use of big data in online education such as privacy, individuality, data security and data ownership (Chen & Liu, 2015; Johnson, 2014; Prinsloo & Slade, 2017; Reidenberg & Schaub, 2018; Williamson, 2017b). Other than work done by Williamson (2017a; 2017b; 2021), there is a disparity between the most prominent and recent works from the first group and the ones from the second. Namely, whereas works in the first group deal with issues relating to economic fairness and equity, most of the work done on big data and online education deals with ethical issues such as privacy or data ownership and security. Therefore, there is a lack academic literature relating to economic fairness and digital labour in light of the ethical issues concerning the use of big data in online education.

Consequently, the primary area of study before data collection and analysis, critically addresses big data, not only as a technology but as part of an economic logic that shapes the social and economic relationships in online education. This primary conception guided the initial steps for data collection and analysis.

Once these initial steps were taken, a more fluid form of the research problems started to emerge from the data, changing as the study progressed. Firstly, issues regarding the sheer amount and variety of information that needed to be processed in order to clearly understand the use of big data started to emerge. As the data collection and analysis process progressed, and categories started to emerge from the data, gaps in the data started to surface, and the data sources and samples had to be broadened in order to fill in those gaps and progress from coded categories towards more comprehensive concepts. For example, this signified incorporating the privacy policy, and terms and conditions documents of third-party partners of Coursera and Blackboard, as an additional data source. Lastly, once the core category, *exploitation of the learning community*, emerged from the data, the final problem appeared to be the lack of conceptual clarity. Meaning, the lack of explanatory power to holistically relate the concepts to the core category, and explain how the core category is supported, or maintained by the concepts emerging from the data. To tackle this issue, a new research question was added to the thesis. This will be discussed in the next section.

Research Aim and Research Questions

By employing a qualitative, critical grounded theory methodology, and engaging in a theory building process, with this thesis I am particularly concerned with contributing to the current body of literature that deals with the ethical concerns of big data in online education. More precisely, I intend my contribution to be of theoretical or conceptual nature. As the study examines only two specific cases of the use of big data in online education, it does not aim at developing a generalizable, or entirely applicable theory that fully explains and predicts phenomena, but rather a theory that helps in the study and interpretation of social phenomena related to the use of big data in online education. Therefore, the preliminary aim of this study was to bring about greater conceptual clarity or a conceptual framework particularly concerned with the ethics of big data practices in education. Following the primary data collection and analysis, the aim of the study got narrowed down to specifically looking into the conceptualization of issues relating to the use of big data, and economic fairness in e-learning.

In contrast to many other research methodologies where the research questions are strictly defined prior to the beginning of the study and they guide the research, in GTM the research questions emerge from, and are refined by the data (Curcliffe, 2005). Furthermore, in line with grounded theory (GT) principles, the format of the research questions in this thesis are open, exploratory, and do not aim at verifying hypotheses, but creating them.

In order to stay open to emergent questions and concepts from the data, I chose to only pose one preliminary broad question that allowed me to approach the data openly and inquisitively, yet with a clear topic in mind. In the words of Glaser, this preliminary research question let me engage the initial stages of the research with the "abstract wonderment of what is going on" (1992, p.22). The first research question is as follows:

- *How and for what purpose is Big Data used in online education?*

This question allowed me to engage with other emerging questions and problems that came to light throughout the data collection and analysis. Thus, in line with grounded theory principles, new, narrower research questions emerged after the initial data collection and coding. Further, these questions were then refined based on the data, emergent categories from it, and theoretical memoing. From the emergent questions and problems, one central research question was defined:

- *What is the role and impact of Big Data on labour relations and economic fairness in online education?*

This research question is central due to its synergistic relationship with the data and the study. On one hand, it is informed by the data and was arrived at by analysing and 'following' the patterns in the data, and on the other, it served as a guiding tool for further exploration and analysis.

Once, I reached a certain level of theoretical saturation regarding the second research question, I noticed that there were some definite conceptual and explanatory gaps in the emerging theory. More precisely, whereas a conceptualization or a map of the economic model and logic of big data in online education was developed (or discovered), an explanation as to why and how is that model maintained, was missing. This led to the emergence of a new research problem and question:

- *How is the economic model of big data in online education maintained?*

As the research progressed, and new problems and research questions emerged from the data, the study adopted new, additional aims as well. More specifically, the pursuit to critically address economic fairness and labour relations in online education, in light of the big data revolution.

Later on, in Chapter 4, in the Data Analysis section, I explain in more detail how these emerging aims and research questions were shaped by the categories and patterns in the data.

Significance of the Study

The significance of this thesis is situated in the need to address the problems that emerged from the data analysis, the gap in the literature, the two cases being investigated, and finally the potential benefit to the online learning community, especially learners and teachers.

Conveniently, for me as a researcher, the problems that emerged from the data, also coincided with the gaps in the literature. Potentially, because the literature was also treated as a data source and integrated into the study.

In the current literature, multiple works have raised ethical issues regarding the use of big data in online education. For example, a number of scholars have researched and brought to light concerns over privacy and consent (Fischer et al., 2020; Reidenberg & Schaub, 2018; Wang, 2016), data security (Kalota, 2015; Raitman et al., 2005), data ownership (Amirault, 2019; Lynch, 2017). Furthermore, a lesser number of scholars mention and address the exploitation of learners for commercial purposes (Marshall, 2014; Williamson, 2021). Lastly, an even more limited number of scholars combine these multiple concerns to form a holistic view of the ethical concerns of big data use in online education and provide a conceptualisation on the matter (Williamson, 2017a). The significance of this study lies exactly in this shortage of conceptualisation, and the lack of conceptual mapping and clarity.

Moreover, the matter of big data ethics in online education is rarely addressed by studying specific cases of e-learning companies and their business practices. Such examples include Williamson's

studying of Pearson's edu-business practices (2021), Vaidhyanathan's work on the 'Googlization' of universities (2009), and Perrotta's et al. of Google's new e-learning platform, Google Classroom (2021). By examining the cases of Coursera and Blackboard, to my knowledge, this will be the first dual case study that takes a critical approach to the use of big data on e-learning platforms.

Lastly, and most importantly, I hope that this study will be most significant and beneficial to the people and organizations that are exploited by the commercial use of big data in online education, the learning community of teachers, students, independent learners and other practitioners. Taking a critically realist ontological stance, I hope that this thesis will be able to uncover and bring to light some previously unquestioned mechanisms and practices, and spark conversations that would eventually lead to change.

Structure and Summary of the Thesis

The thesis is divided into six chapters, and each chapter consists of multiple smaller sections. In this section, I will present the structure of the thesis and summarize each of the chapters.

Chapter 1. In Chapter 1, the topic of the thesis is first introduced. Then, the research problem, aims, questions, and rationale are presented. The chapter finishes by providing a structure and a summary of the chapters in the thesis.

Chapter 2. Chapter 2 consists of three parts. The first part is a brief explanation of the role of a literature review in a GT study. The second part attempts at defining 'big data' as a key term in the scope of this study. Lastly, the third part is a short review of the literature that is supposed to serve as an initial foundation for the research.

Chapter 3. In Chapter 3 the research methodology of choice, Critical Grounded Theory, is discussed.

Chapter 4. Chapter 4 deals with the methods of the research. Namely, it starts by laying out the sampling strategy and rationale. Then, it proceeds to discuss the different data sources and data collection processes. Finally, it presents and discusses the data analysis methods.

Chapter 5. The fifth chapter presents the findings of the research, as well as an integration of the findings with the existing literature, by presenting the core category and summarising the emergent theory. Furthermore, the summary is followed by an evaluation of the theory.

Chapter 6. The last chapter starts by discussing the implications of the findings and the emergent theory. Consequently, the limitations of the study are discussed, and suggestions for further research are given. The chapter and the thesis ends with a concluding section.

Chapter 2: Foundational Literature Review

Grounded Theory and Literature Reviews

The use of a literature review in GTM is probably one of the most controversial points regarding this approach. The contention stems from the classical grounded theorists' view that the researcher should not engage with the literature before conducting some initial data collection and analysis (Glaser & Straus, 1967; Glaser, 1998; Glaser and Holton, 2004). This delay in conducting the literature review in the initial stages of the research is meant to prevent the researcher from contaminating the data collection and analysis process by imposing existing theories and knowledge. Hence, some researchers believe that in GT one should commence the research as a "blank slate", without any prior knowledge or agenda (Suddaby, 2006). Thus, avoiding the literature review at the beginning of the study.

However, multiple subsequent scholars and grounded theorists have challenged this view, rejecting the idea of the researcher as a "blank slate" (Suddaby, 2006; Timonen et al., 2018; Urquhart & Fernandez, 2016). In fact, numerous scholars have recognized the need for some prior knowledge on the topic of research before commencing the study. Ignoring the literature on any prior empirical and theoretical knowledge on the topic is not only unexpected from the researcher but also impractical and unnecessary (Goulding, 2002; Suddaby, 2006, Timonen et al., 2018). Both Timonen et al. (2018), and Andrew (2006) argue that the key to a successful GTM study is for the researcher to remain theoretically sensitive, and open to the discovery of any, even unanticipated concepts, patterns and interpretations from the data. On that note, Andrew argues that there may be two different literature

reviews that can be conducted in grounded theory, the preliminary and the integrated one (2006). The preliminary literature review or the foundational literature review is the "one that puts the study into some context". The second literature review, or the integrated one, takes the literature as data and is used to integrate the emergent theory with the rest of the work in the field of study. The latter one, in this thesis, is presented in the fifth chapter, concurrently with the discussion of the findings.

The role of this foundational literature review is multi-faceted. Firstly, it identifies the preliminary area of concern for the research, puts the study into context, informs the rationale of the study and justifies the research questions (Andrew, 2006; Coyne & Cowley, 2006). Secondly, it provides both the researcher and the reader with an awareness of the existing theories, and attempts at defining key concepts in order to more productively "remain open to the portrayals of the world" (Timonen et al., 2018, p.4). Thirdly, it stimulates the 'theoretical sensitivity' of the researcher (McGhee et al., 2007). Meaning, it enhances the ability to perceive concepts and the relationships between them in order to allow for theory building and conceptualization in the later stages of the research.

The role of this literature review however is not to inform of, or assist with the emergence of a theory or a core category. That is the role of only the primary empirical data gathered through data collection. Further, the literature review should not provide hypotheses that the researcher verifies. Lastly, the foundational literature review does inform of pertinent knowledge lacunae in previous research or the literature. However, these gaps in the knowledge represent a mere direction for the initial steps of data collection and analysis. When gaps in the literature are identified, the researcher runs the risk of forcibly fitting the data in these gaps and creating a preconceived idea of what the data should address (Glaser, 1978; Glaser, 1992). Furthermore, the gaps in the literature will be integrated with the emerging gaps, problems, and concepts from the data in the latter literature review.

In order not to force the data into theoretical assumptions encountered in the literature, and force the researcher into testing hypotheses, the researcher should be self-aware and cautious of how the literature can influence the trajectory of the research. The foundational literature review should be quite broad in focus, not merely concentrating on one substantive area, but on a broad spectrum of seemingly unrelated or opposing literature (Dunne, 2011; Suddaby, 2006). Therefore, for this

literature review, I will be covering literature dealing with big data both in the field of online education and as a whole, in wider socio-technical settings. Moreover, I will also include works focusing both on the benefits and the concerns regarding big data in online education. Nevertheless, before commencing the initial literature review, I will attempt to define big data as a key term for this thesis.

Key Terms and Concepts: Big Data

In the past two decades, big data has increasingly become a loaded term with many preconceptions tied to it. Often, people have polarizing attitudes towards it. For example, many companies see it as a marketing tool, or a buzzword by advertising their 'data-driven' solutions, or 'powered by data' products. On the other hand, big data is often associated with negative allusions to social control, surveillance, lack of privacy etc. For instance, in light of the recent Cambridge Analytica data scandal, many media outlets came out with headlines such as: "Big Data is Watching You!", alluding to Orwell's *1984*, a dystopian fiction novel dealing with themes of mass surveillance, totalitarianism, and strict social control (Bartlett, 2018).

However, despite the popularity and wide use of the term, there is a clear lack of understanding of what big data really is. To truly understand big data, we first need to look at the technical definition. In technical terms, big data seems to be differentiated from 'small data', or statistics and analytics. Many scholars, including scholars in the field of education, take the three Vs as the defining features of big data (Fischer et al, 2020; Kalota, 2015; Lee, 2017; Wang, 2016). These three Vs relate to the volume, variety, and velocity as the three defining dimensions of big data (Laney, 2001). Nonetheless, recently, these three dimensions have been expanded to five, adding value and veracity to the list (Emmanuel & Stainer, 2016). In short, big data is defined by the fact that we are generating, collecting and analysing record-high amounts of data, at an ever-faster rate, from a diverse set of sources. Furthermore, this data is reliable, accurate, and provides some commercial or scientific value (Lynch, 2017). Multiple other characteristics, such as scalability, or high indexicality, have been also added to these dimensions, all of which contrast previous data collection and analysis practices. Despite the neatly defined technical qualities of big data, Kitchin and McArdle have found that only a

few proclaimed ‘big data’ systems hold the technical qualities defined in the literature (2015). In fact, they state that the 3 Vs are unnecessary when trying to define big data and that rather, we should focus on the ontological framing of big data. They conclude that there are many “species” of big data, each with its own defining attributes, and they call for a shift away from defining big data in “generalities”, towards using specific qualities that will lead to more clarity (Kitchin & McArdle, 2015).

Consequently, to define big data for the purpose of this paper, we ought to not only take the general technical qualities of big data present in the literature into account, but also how big data operates in the field of online education, the logic that underpins it, and its societal implications. After all, the big data used to predict earthquakes in Chile is, and should be, different compared to the big data used to predict whether a student will enrol in a specific online course.

Firstly, it is essential to mention that when defining big data in online education, we are not completely separating it from the generalities such as the 5 Vs, in fact, these technical qualities largely apply to the big data used in the field. However, as Williamson has already argued, instead of reducing big data to technical characteristics and qualifications, we must think of it as a “sociotechnical system” (2017a, p.65). Perhaps, to give an account of big data in online education there are two domains one must address: the technical and the social.

Whatever definition we give to big data, there are few indispensable functional components that make big data work. Firstly, digital data needs to be extracted and stored. Once collected, this data needs to be made sense of, or in other words, it needs to be analysed. For this purpose, we need algorithms, mathematical and statistical models that are run on powerful computers that can process large amounts of data. These models through pattern recognition ‘learn’ from the data, and are then able to cast accurate predictions about a new set of data (e.g. what course a student might be interested in taking next), or group the data in practical classifications (e.g. classifying students based on the risk of dropping out). Since the models ‘learn’ from the data, the data that they learn from is usually called ‘training data’, and often the algorithms used are a product of a field of study called machine learning (Kapitanova & Son, 2012). In more emblematic, logistical terms, big data is the process where the

raw materials are large training data sets. Then, these data sets are worked, or analysed by computers with the help of tools called algorithms. Lastly, the final product is a prediction or a classification of new data of a similar kind to the training one.

Williamson argued that the nature of big data and the approach one should be taking when defining it is, and should be, sociotechnical (2017a). Meaning that the technical aspects of big data are closely linked with the social implications and practices surrounding it, and they are in a reciprocal relationship where one shapes the other and vice-versa.

The underpinning social logic of big data in education can be explained by defining the producers and the subjects of the data, the owners and beneficiaries of the process, and the incentives and social outcomes. Data in online education is primarily extracted from platforms that are mainly used by learners, teachers, and administrators. Therefore, most of the data in online education comes from the experiences and actions of these groups. Additionally, the productised version of the data is also then used on these platforms, meaning that the groups that use the platforms most are the ones that are most affected by big data. This makes learners, teachers and administrators both the producers and the subjects of big data in online education (Finn, 2016).

Furthermore, the owners of the data are those who extract it and have the economic power to process it and analyse it. In the case of online education, these are large educational companies, online education providers, and educational institutions such as universities. As the owners and controllers of the data, they are also the ones that are set to gain the most from it. However, in exchange for more data, learners, teachers and administrators might benefit in non-monetary terms, such as better service, increased success, or improved learning experience.

Whereas the producers/subjects of the data are partaking in the process of big data because that's the dominant paradigm of practice in online education, the owners and controllers of the data have commercial incentives. With commercial incentives in mind, the companies and institutions are interested in extracting more data. This unprecedented scenario has led to some serious and controversial debates regarding the social consequences of big data. Particularly, concerns have been

raised regarding the excessive tracking, monitoring and even altering of individual and group behaviour and action, privacy, and the commodification of human experience (Williamson, 2017a).

For the sake of simplicity, I have tried to briefly define big data in both technical and societally terms. However, any attempt to clearly and holistically conceptualize big data in online education in such a small space is doomed to over-simplifications and oversight. As this task is outside of the scope of this study, more extensive work is needed to fully understand and conceptualize big data in online education.

Foundational Review of the Literature

The purpose of this literature review is to put the study into context, provide awareness and knowledge of existing theories regarding the topic, and support and foster 'theoretical sensitivity'. For this purpose, the guiding beacons of the foundational review are the preliminary aim and research question stated in Chapter 1. The intention for this literature review is not for it to be an exhaustive review, but rather an appraisal of a broad, and diverse set of work on big data. In order to remain broad in focus yet stay informative, I employed strict search methods and strategies, and used specific key terms.

Big data and the Internet

Since the conception of the internet, the amount of data generated, consumed and used by humans has exponentially grown. In fact, from the beginning of recorded human history until the early years of the internet in 2003, humans have generated a total of five exabytes of information. However, as the internet grew in population and size, the numbers of generated data grew with it. For example, in 2011, humans were creating five exabytes of data every two days. This number rose to five exabytes every 10 minutes in 2013 and continues to grow exponentially (Zwitter, 2014). The internet and big data are closely linked. Precisely, the internet is where big data happens, and big data is a crucial part of how the internet works. Moreover, even though enabled and run by technology, the internet is primarily a social phenomenon and a defining part of human communication in the 21st century (Memmi, 2015).

Thus, as a concept, big data is relevant to many different technical and social disciplines, from information system studies to economics. Each of these disciplines uniquely approach big data, providing perspectives and views that are overlooked by others. However, none of them neglects or questions the socio-technical nature of big data and its central place in the techno-cultural arena of the internet. In order to gain broad awareness about big data, in this section of the literature review, I will examine and present different topical perceptions and insights from a diverse set of disciplines including information systems and computer science, sociology, ethics and philosophy, and economics.

Firstly, in the field of information systems, even though a great amount of discussion and work is dedicated to the technical part of big data analytics, some attention has been paid to the value that lies in big data (Chiang et al., 2018; Günther et al., 2017; Loebbecke & Picot, 2015). Günther et al., categorizes value realized through big data in two: social value and economic value (2017). On the one hand, social value pertains to the improvement in social wellbeing, particularly in fields such as education and healthcare. On the other, economic value entails an organization's increase in monetary gain or a competitive advantage. This can include value in the form of product development, customer behaviour research, operational and strategic decision making, and many more (Günther et al., 2017). This stratification of the applications and value realization of big data is further evident in Chong & Shi (2015). They divide applications of big data into three domains. Similarly to Günther et al. (2017), they start by differentiating between the business and social applications of big data. However, Chong & Shi, also add the scientific application of big data, arguing that contemporary scientific research is heavily reliant on the value nested in big data sets (2015).

Furthermore, work on the ethical concerns and the interests of often marginalized stakeholders such as the users and customers have been raised (Günther et al., 2017, Ekbia et al., 2015; Kennedy & Moss, 2015; Newell & Marabelli, 2015;). Interestingly, in most of the literature in the field of information systems, these ethical concerns are presented in opposing dualities. On the one hand, we have the legitimate concern of the users and customers, and on the other, a counter-argumentative response to these concerns is presented. For instance, Ekbia et al. phrase the issue of opening data to be freely

available to users and consumers as a question "*To open or to hoard?"* (2015). Similarly, when addressing the issue Günther et al., presents either an open or a controlled approach to big data access, providing arguments for both (2017). This is best illustrated by Newell & Marabelli's depiction of "trade-offs" on societal issues that are associated with big data (2015). Such trade-offs include "privacy vs. security" or "control vs. freedom". An example of this trade-off can be understood through a portrayal of how users trade their freedom of choice, for the sake of controlled convenience through recommendation systems or personalized content distribution practices (Newell & Marabelli, 2015).

Contrastingly, the academic literature in the field of big data ethics, presents a more critical view of big data practices, arguing that the increased adoption of big data has outpaced awareness, as well as concerns regarding transparency, privacy, openness, wealth and power distribution, propensity, consent and ownership (Jurkiewicz, 2018; Richards & King, 2014; Richterich, 2018; Zwitter, 2014). For instance, unlike some of the above-mentioned works in the field of information systems, the literature on big data ethics does not differentiate between the social and economic value extraction and generation of big data. In fact, Jurkiewicz raises the issue of data collection and extraction "under the guise of social betterment" (2018, p.48).

Privacy is one of the most extensively mentioned and elaborated on ethical concerns in the literature. This focus on privacy stems from the fact that the early discourse regarding ethics in information technology mostly revolved around the value of privacy (Regan, 2000), and that in many legal frameworks and traditions, especially in the United States, privacy is considered a fundamental civic right (Glenn, 2003). Furthermore, privacy is considered as an umbrella term for many other ethical issues related to data such as security, surveillance, ownership, anonymity and others (Regan & Jesse, 2019; Richterich, 2018). Due to the commodification of personal data, some scholars connect the issues of privacy with concerns regarding a new economic reality (Wielki, 2015).

Zwitter separated the stakeholders in this new economic reality into big data collectors, utilizers, and generators (2014). With the collectors and utilizers wielding most of the power and wealth. In fact,

much of the critical literature on ethics and data economics discovers that the role of the generators is to perform free labour and produce free data by consuming and creating on digital platforms (Ritzer & Jurgenson, 2010; Scholz, 2012).

The ethical concerns regarding the distribution of wealth and power are even further expanded on in the works of scholars of economics who try to conceptualise capitalism in the new information era. Whether it is in surveillance capitalism by Zuboff (2019), the Digital/Big Data Capitalism of Fuchs and Chandler (2019), or in Srnicek's Platform Capitalism (2017a), big data plays a central role in the economics of the internet. All three of these conceptualisations claim that the economic institutions of capitalism have undergone a major change, with data as a central resource (Fuchs & Chandler, 2019; Marciano et al., 2020; Srnicek, 2017b; Zuboff, 2019). Practically, all of them are explaining the same phenomenon, using a similar critical framework, and come to somewhat similar conclusions. For instance, they all agree how this new form of capitalism is detrimental to the social fabric of any democratic society by concentrating power, wealth, and knowledge. Zuboff argues that surveillance capitalism is antithetical to democracy, and is a form of "tyranny that feeds on people, but it's not of people" (2019, p.479). Similarly, Srnicek argues how the emergence of large platform capitalist companies are a serious concern as they "are becoming owners of the infrastructures of society" (2017a, p.62). Lastly, Fuchs and Chandler, state that digital capitalism's "structures of domination and exploitation threaten social cohesion and democracy" (2019, p.3).

Some other common themes between these three approaches when defining the economic milieu in the era of big data include the extraction and use of behavioural data for profit, the network effect of big data and its monopolizing nature, and the appetite of big data companies for more data. However, there are some differences present, but they are mostly differences in focus, rather than content and context. In other words, the difference between these three approaches is what they focus on when conceptualising capitalism in the big data era, rather than what their conclusions are. For example, whereas Zuboff (2019, p.14) focuses on the behavioural surplus generated through data practices and the criticism of the commodification of "human experience", Fuchs specifically focuses on labour relations and class struggle in digital capitalism (Fuchs, 2012; Fuchs & Chandler, 2019).

Nevertheless, between the literature on big data ethics and big data economies, there is a common notion that the economic, technical, and social realities of big data are not separate, but rather interconnected and inseparable. Thus, in a field such as education, that has been labelled as a domain of social, rather than an economic value for big data (Günther et al., 2017), one must critically engage with the economic incentives of key actors.

The Story of Big Data in Online Education - Benefits and Concerns

The story of big data in online education is closely linked with that of Learning Analytics (LA). Learning analytics is the practice of collecting, measuring, analysing and reporting educational data, in order to improve the understanding of learning and teaching processes, and tailor and personalise education for each student (Johnson et al., 2011). With the mass adoption of online and blended learning and the move of vast amounts of educational activities online, big data has taken a central role in the field of education (Johnson et al., 2013; Seufert et al., 2019). This move to online learning entails that learners and teachers increasingly perform more and more activities on educational platforms. These activities generate data that can provide lucrative insights into the learning process, performance, and activities of learners (Elia et al., 2019). In fact, much of the literature on big data in online education points to the "goldmine" of unused learning data that is just waiting to be discovered and utilized by academic institutions and companies (Drigas & Leliopoulos, 2014; Romero & Ventura, 2017).

Furthermore, there is a thread regarding the revolutionizing and game-changing potential of big data in the online education literature (Drigas & Leliopoulos, 2014; Gibson, 2017; Kalota, 2015; Reyes, 2015). The applications of big data in education are vast and various. Some of the applications include improving student success and predicting and decreasing dropout rates (Brown, 2011; Greller & Drachsler, 2012; Kalota, 2015; Liang et al., 2016); providing a personalised learning experience through data-driven recommendations and suggestions (Dishon, 2017; Siemens et al., 2012; Verbert et al., 2012); assistance with grading and assessment (Lynch, 2017; Mitros et al., 2013); and measuring and predicting student satisfaction (Elia et al., 2019). However, these benefits do not come without challenges. For instance, Menon et al. (2017), argue that big data practices in education are yet to be

successfully used to their full potential. Some common issues include the complexity and treatment of data (Marín-Marín et al., 2019); the diverse sources and datasets and the failure to combine them or join them (Reyes, 2015); and technical issues regarding the synchronization of different types of data (Geller & Drachsler, 2012).

Besides these benefits and challenges, there are also ethical concerns taken into account regarding the use of big data in online education. The main ethical issue that is raised by the bulk of the literature is privacy. Reyes takes a more pragmatic approach arguing that if the learners suspect that their privacy is being invaded, they might not be willing to share any data (2015). Pardos raises concerns regarding the vast collection of behavioural data and its effect on privacy protection, however, they call for a cost-benefit analysis when it comes to dealing with privacy-sensitive data (2017). Marshall (2013) raises similar concerns. Lastly, Reidnberg and Schaub, go more in-depth into privacy in online education and the role of big data in the issue, and provide policy recommendations to move past this concern (2018). Moreover, other commonly raised issues include data security (Kalota, 2015; Reyes, 2015); data ownership (Amirault, 2019; Chen & Liu, 2015; Johnson, 2014; Lynch, 2017; Prinsloo & Slade, 2017); surveillance (Regan & Jesse, 2018); and individuality (Johnson, 2014). Furthermore, the issue of economic exploitation and wealth concentration is rarely present in the literature and only on the margins. Marshall (2014), for instance, raises questions regarding the commercial exploitation of learners on various MOOCs platforms. Most extensive work on this topic has been done by Ben Williamson, providing an account of the ethical consequences of the datafication of education (2017a), the concentration of power when it comes to analysing big data in online education (2017b), and a case of a platform edu-business, alluding to the before-mentioned platform capitalism (2021). Nevertheless, Williamson's work does not provide a more holistic and conceptually clear view of the underlying economic model behind the datafication of education.

Key takeaways

Some key takeaways from this literature review, as a foundation of this thesis, are:

1. Big data as a concept is constantly evolving, and rather than a technology, one can see it as an unprecedented economic and societal logic with unique ethical concerns regarding privacy, economic exploitation and concentration of power and knowledge.
2. Economic fairness and exploitation are central themes in the wider literature on big data ethics, however, in the field of big data and education, these themes are marginal and rarely addressed.
3. Whether it is regarding big data as a term or the underlying socio-economic logic that big data operates under within education, there is a lack of conceptual clarity in the field.

Chapter 3: Research Methodology – Critical Grounded Theory

Researcher's Position

The idea of the researcher as a "blank slate" in grounded theory is unrealistic, unproductive and misleading (Suddaby, 2006; Timonen et al., 2018; Urquhart & Fernandez, 2016). Many grounded theorists argue that the 'guiding interests' of the researcher can serve as a 'point of departure' for the grounded theory study (Charmaz, 2006, p.17). However, it is important for the researcher to be transparent by clearly stating their starting positions, and stay open-minded by expecting and allowing for these starting positions to change as data is collected and analysed. Being transparent is important since "the more we are aware of the subjectivity involved in data analysis, the more likely we are to see how we are influencing interpretations." (Corbin & Strauss, 2008, p.33).

Therefore, acknowledging my active and subjective involvement in this research, I will list some of my 'points of departure' for this thesis. Firstly, my ontological positioning lies under the umbrella of critical realism. Meaning, that I do not merge reality with the knowledge we have and can have about reality. Reality exists independent of our knowledge of it, however, the descriptions of this reality are mediated by human activities and practices, such as social contexts or language (Oliver, 2012). Therefore, due to our inability to describe reality, in reality's terms, our knowledge will always be

limited to our perspectives of it. As a realist researcher, my goal is to produce knowledge that can generate truer or closer descriptions and explanations of reality. Moreover, with this work, I aim at uncovering processes, mechanisms, and structures that are hidden from perception but generate empirically observable social relationships. Thus, in essence, my goal as a researcher is to produce emancipatory knowledge that fosters change.

Secondly, as a researcher, I am interested in the space where technology and education meet. More precisely, I am particularly interested in the ethical implications of the use of technology in education. Artificial Intelligence and big data have been my particular areas of interest in the past two years. Drawing on my observations, previous knowledge, and experience I have come to realize the unprecedented nature of big data, not only as a technology, but a social force and logic that dictates social and economic inequalities on the internet. Therefore, I approach this study from a position of criticality and concern.

Choosing the Right Methodological Approach

By aiming to bring about conceptual clarity, this thesis warrants a methodological approach that is suitable for theory building or development, rather than empirical theory testing. Furthermore, in this thesis, I am not trying to apply the data to an existing theory, but rather arrive at a conceptualisation or a theory through and from within the data. Furthermore, the fact that the topic of big data in online education has seldom been critically examined requires a methodology appropriate for explorative research into a topic that has not been studied regularly. Due to these three reasons, I considered grounded theory as the appropriate research methodology for this thesis. Firstly, grounded theory as a method seeks to generate new explanatory theories (Corbin & Strauss, 2008). Secondly, the essence of grounded theory is that the theory building process is grounded in the data, and hypothesis testing is avoided (Suddaby, 2006). Lastly, grounded theory is specifically appropriate for "discovery-oriented" research in areas of study that are under-theorized (Burck, 2005, p.244). Consequently, for this thesis, I will adopt the principles, practices, and guidelines of grounded theory in order to conduct the data collection, analysis, and theory building.

Grounded Theory

Grounded Theory as a methodology was first used by Glaser and Straus (1965) to research terminally ill patients and their perceptions on dying. Later on, in 1967, in the *Discovery of Grounded Theory,* they expanded upon and conceptualised grounded theory as a research methodology. At the time, the dominant paradigm in sociological research was positivism. Meaning, that most of the research was focused on testing and verifying existing theories, which led to a lack of inductive theory building (Chamberlain-Salaun et al., 2013). Displeased with this lack of theory building, and the over-reliance on positivism, Glaser and Straus created grounded theory as an alternative to the positivist paradigm, providing a rigorous qualitative research methodology with set ontological and epistemological positioning fit for theory development (1967). As such, by exposing the limitations of the dominant positivist approaches when trying to generate novel explanations for social phenomena, grounded theory presented an important critique to the positivist research approaches (Denzin & Lincoln, 2008). Glaser and Strauss challenged the dominant positivist stance on how research should be done, by introducing two fundamental concepts of grounded theory, constant comparison and theoretical sampling (1967). The concept of constant comparison challenged the positivist idea that data collection and analysis should be a linear process with clear boundaries. In grounded theory data analysis starts at the moment when the first bit of data is collected, and it is an ongoing cyclical process that leads to theory building. Secondly, unlike the dominant modus operandi, in which data collection was dictated by previously set hypotheses, in grounded theory, by the way of theoretical sampling, data collection is determined and constantly updated by the emerging conceptual categories from the data (Suddaby, 2006).

Since its initial introduction, even between the original authors, there have been multiple points of contention on how grounded theory should be done. Therefore, multiple branches of GTM often are in dispute with one another. The first form of GT is the one closest to the initial methodology presented in 1967. This theory is often named Classical Grounded Theory. Ontologically, due to its objectivist underpinnings, classical grounded theory is closest to positivism and generally is considered to be post-positivist in nature. In Classical grounded theory there is an objective theory to be discovered

from within the data and regardless of who the researcher is, if the methodological procedures are correctly followed, the same theory will emerge from the data time and time again (Glaser & Holton, 2004). The Classical model mostly puts emphasis on the inductive nature of GT research and the importance of limiting all biases and preconceptions of the researcher. In order to achieve this objectivity, Glaser argues for a broad research problem and question, and very limited, or even no literature review before starting the data collection and analysis process (1992).

The second variant of GT, known as the Straussian model, was first introduced by Corbin and Strauss (2008). Similarly to the Classical variant, the Straussian model is ontologically situated in post-positivism. However, there are some epistemological differences between the two. Namely, Strauss and Corbin propose a data analysis process that includes both induction and deduction, and propose a method of ongoing validation (Timonen et al., 2018). Glaser, criticizes this model, arguing that it has diverged from GTM, and the introduction of validation to the methodology, defeats the purpose of the original idea (1992). The Straussian model also is not opposed to including the literature and the professional experience of the researcher, however, they must be used as additional, supporting data (Corbin & Strauss, 2008).

In response to the ontological homogeneity between these first two variants, Charmaz (2000) developed the Constructivist grounded theory. For the Constructivist approach, a central feature is the recognition of subjectivity, and the active role of the researcher in the data collection, knowledge creation, and theory building process (Charmaz, 2006). In Constructivist Grounded Theory, the researcher is not an objective and detached observer, but rather a crucial participant in the data generation and theory building processes (Timonen et al., 2018).

One of the latest variants of grounded theory is that of Critical Grounded Theory (CGT). Ontologically, divergent from both the constructivist and post-positivist positions of the previously mentioned branches, CGT is aligned with Critical Realism (Looker et al., 2021). Furthermore, CGT introduces retroduction as a data analysis tool and a form of critical inquiry (Timonen et al., 2018). Lastly, different from the previously mentioned approaches, the researcher in CGT starts by conducting critical observations and seeks to enact change, specifically seeking to produce

emancipatory knowledge relating to power dynamics, equality and social justice. Since this is the variant most relevant to this thesis, I will elaborate on it in more detail in the next section.

Critical Grounded Theory

Critical grounded theory (CGT) is divergent from all the other variants of GT in three aspects. Firstly, ontologically it is neither based on the post-positivist nor the constructivist paradigm, rather it aligns with the critical realist ontology. Secondly, it is concerned with creating critical, emancipatory knowledge regarding issues such as power, justice, and equality. Thirdly, it introduces retroduction as a mode of critical inquiry. In this section, by addressing all of these three points, I will attempt to explain CGT as the appropriate grounded theory variant for this research study.

One of the crucial features of critical realism is the separation between knowledge and being, or simply reality and our knowledge of it. Bhaskar, the founder of critical realism, calls this the epistemic fallacy (1978). Bhaskar rejects the idea that reality is conflated with our knowledge of it, therefore, accepting the positivist claim of a reality outside of human conception, but criticizing the positivist assumption that ontological questions can be answered epistemologically (1978). The relationship between the real and knowledge, and the ontological and epistemological is further explained by the stratification of reality into three; the real, the actual, and the empirical (Looker et al., 2021). The real is the realm of intransitive mechanisms and structures, and these structures are not influenced by our knowledge or experience of them (Bhaskar, 1987). The domain of the real is based in ontology, rather than epistemology. Everything that happens in the actual domain, all the events and phenomena, are caused by mechanisms in the real domain. Thus, there is a causal relationship between the real and the actual. In other words, the actual domain is a manifestation of the real that we may or may not observe. Lastly, the empirical domain, the domain of epistemology, is where empirical data in the form of action or experience can be observed (Looker et al., 2021). Unlike positivism, critical realism does not claim that one can arrive at knowledge about the real, which is ontological, by engaging with the empirical, which is epistemological. On the other hand, unlike constructivist philosophies, critical realism argues that our knowledge, interpretation and descriptions

that stem from the empirical have no impact on the real. The real domain is intransitive and enduring, whereas the empirical is transitive and changing (Dobson, 2002).

The reason why critical realism is ontologically attractive is that it combines elements of positivism and constructionism to produce a philosophy that bridges the divide between the two (Taylore and White, 2001). It admits that even though there is the need for seeking evidence of a reality external to human consciousness, any meaning that we make of that reality is, and will be socially constructed (Oliver, 2012). Thus, although in the real domain there is knowledge that is objective and devoid of human interpretation, we can only explain and communicate this knowledge in empirical terms, which are open to interpretation (Looker et al., 2021). Further, the actual domain allows the creation of theoretical explanations that are not empirically evident or observable. The critical realist bridge between constructivism and positivism can truly be observed in the argument that events and experiences observed in the empirical domain, may or may not be affected by unobservable theoretical constructs of the actual domain (Brown et al., 2002; Looker et al., 2021).

The emancipatory objective of critical realism and CGT, and the drive to enact change through research is central to this thesis. By critically examining the generative mechanisms of observable events and experiences, critical realism provides a framework for uncovering, explaining, and therefore altering hidden social structures that may have an impact on human social wellbeing (Oliver, 2012). Bhaskar explains this emancipatory goal of critical realism as the need to move people away from a demi-reality, which contains oppression, exploitation, and alienation (2002). Some grounded theorists, especially those adhering to Classical grounded theory, might be sceptical of initialising a study, choosing research problems and questions with an explicit moral and societal goal in mind. Firstly, the researcher runs the risk of explicit data forcing, and secondly, it introduces the possibility of biased theorisation (Hadley, 2019). However, Hadley (2019), further argues that if the researcher approaches the study with an open mind, and the research is done in a transparent, honest and reflexive manner, there is no reason to believe that the critical grounded theorist forces the data into already existing theoretical assumptions.

In CGT, one of the goals of the researcher is to produce a theory that describes the empirical domain of the study by defining and explaining the mechanisms that generated those empirical observations. These generative mechanisms often derive from the real domain. In order to search for the *real* mechanisms, CGT introduces one of the key tools for a critical realist inquiry, retroduction (Looker et al., 2021). Put simply, retroduction is the practice of asking "what must be true for this to be the case?" during the data analysis process (Oliver, 2012, p.379). Retroduction requires the researcher to constantly oscillate between theory and evidence, and the phenomenon to move from the real to the empirical and vice versa. The goal is to "take the data backwards" through the emerging categories and concepts in order to identify the causal mechanisms of the observed phenomena in the empirical.

Grounded Theory Principles

The following section will present some of the central GT principles used in this study.

Openness

All GTM approaches strive to remain open to new findings, and remaining open to the data is elementary for GT research (Timonen et al., 2018). The key principle here is not to force the data into theoretical accounts, avoid hypothesis testing that can 'close' the research, and remain open to unanticipated findings. As mentioned previously, due to the emancipatory and morally motivated nature of CGT, remaining open to alternative possibilities is even more important (Hadley, 2019). Openness can be supported by acquiring theoretical sensitivity (Timonen et al., 2018), constant memoing and reflection, induction and retroduction (Looker et al., 2021; Sbaraini et al., 2011).

Iteration and the Constant Comparative Method

One of the key features of GTM is iteration, or constant comparison, where data collection and analysis are done simultaneously (Suddaby, 2006). This constant comparison allows for the identification and constant alteration of patterns and causal relationships that emerge between codes, categories and concepts (Bitsch, 2005). Different approaches of GT take different steps when it comes to constant comparison, however, there is a common cyclical two-step process that clearly describes the essence of the constant comparative method. Firstly, the data is 'opened up' in order to arrive at

codes and categories that emerge from it. Secondly, by comparing these codes and categories, a connection between them is drawn to create a conceptual framework and arrive at theoretical concepts (Timonen et al., 2018, p.5). Often to these two steps, there is a third step added where the researcher attempts to reduce the number of conceptual categories, called delimiting (Glaser & Strauss, 1967).

In more simple terms, one can understand the constant comparison method as the comparison between new and previous data which is facilitated by rigorous coding. The goal of the constant comparison method is to identify a core category, produce a substantive theory, and reach theoretical saturation.

Theoretical Sampling

The previously mentioned 'new' data that is compared to the preceding one is arrived at through theoretical sampling. Theoretical sampling is a data generation and sampling method central and unique to GTM. In simple terms, theoretical sampling means that what data should be collected is informed by the emerging theory, meaning, by the categories and concepts that are emerging from the existing data (Suddaby, 2006). The idea behind theoretical sampling is that as data analysis progresses, gaps in the current data set emerge, and questions and dilemmas arise. Therefore, the data informs the researchers what they do not know yet, or what issues should be further examined (Charmaz, 2006; Glaser & Strauss, 1967).

Memoing

According to Timonen et al. (2018), memoing is an "invaluable tool" in the theory construction process. Memoing is the practice of writing brief memos, or comments regarding some codes, categories, concepts, events, relationships, thoughts, or questions during the research process. There are several reasons why memoing is such a crucial part of GTM. Firstly, memos are a form of record-keeping, where the researcher's thinking is recorded. This contributes to the openness and transparency of the research (Bryant & Charmaz, 2007). Secondly, memos provide additional argumentation for the theory-building process (Timonen et al., 2018). Thirdly, following Glaser and Strauss (1967) idea that everything is data, memos can serve as data as well, since they are often include the researcher's observations, opinions, and attitude towards the research problem and data.

When memoing, it is important for the researcher to ask questions regarding the data and the theory development process. Specifically, Glaser (1978, p.57), poses the question "What is actually happening in the data?". Furthermore, Hadley (2019, p.21), poses additional questions more tailored towards a Critical grounded theory study.

In the scope of this research, I wrote over 30 theoretical memos. Memos were written in an informal language, and they served me as tools to reflect upon, comment on, and record the theory-building process. I used memos to record ideas regarding the progression of the theory construction, cast doubts or raise potential issues regarding a specific code or category, or simply briefly comment or provide additional information regarding some of the concepts, categories, and codes.

Theoretical Saturation

In many qualitative studies, the idea of saturation is pursued by researchers. This means that the researchers want to get to a point of the study where they are not getting anything new from the data. For instance, in the case of a study based on interviews, saturation is reached when the researcher is not hearing anything new from the participants (Sbaraini, 2011). This idea is called data saturation, and even though theoretical saturation is something different, the two are often confused (Timonen et al., 2018). Whereas the goal of data saturation is to get to a point where no new meaningful data can be present, the goal of theoretical saturation is to exhaust the coding process to the point where no new meaningful codes or categories emerge from the data. In other words, theoretical saturation is reached when even though new data is being collected and analysed, there is a discontinuation of new coding units being generated (Holton, 2007).

Production of a Substantive Theory

The substantive theory is the form in which the results of a GTM study are presented (Bryant & Charmaz, 2007). Often, some grounded theorists use the substantive theory from their study to generate grounded formal theory. (Glaser & Strauss, 1967). Nevertheless, for most GTM studies, the generation of a substantive theory is the aim of the research. However, this study aims at providing conceptual clarity in the sense of generating an explanatory conceptualization by illustrating,

describing, and analysing issues and phenomena that are contextualised and grounded in a specific setting. Therefore, this study does not necessarily aim towards a full substantive theory that is generalizable, but it aims towards bringing conceptual clarity to the use of big data in these two specific contexts, which may manifest itself in the form of a substantive theory.

Methodological Limitations and Challenges of Critical Grounded Theory for this work

No methodology is perfect and all methodologies present some challenges to the researcher. During this research, I came across several methodological limitations and challenges of CGT. In this section, I will point out the most relevant limitation that is specific to CGT, and the two most pressing methodological challenges for this work.

Firstly, the combination of a case study approach, critical realist ontology, and GTM implies that there will be limitations to the generalisability of the study. For instance, Kempster & Perry (2011) argue that since in critical realism no two contexts and settings are the same, explanatory statements do not seek to produce generalisations beyond the studied phenomenon in the specific context. When combined with the case study approach and theoretical saturation principles followed in this study, where a deep examination is conducted on two specific cases, I run the risk of over-interpretation and being overly focused on the site-specific phenomena and effects. Due to this site-specific substantiveness, as a researcher, I might fail to see the bigger picture and acquire certain critical knowledge (Kempster & Perry, 2011). With that in mind, this thesis is not aiming towards generalisability, but towards the prospect of applicability and comparability to other, similar contexts.

Secondly, balancing between conducting a proper literature review and introducing preconceived notions was a challenge. Whereas Glaser (1978) mentions that the researcher needs to enter that data collection and analysis process with no preconceived notions, many other grounded theorists reject this idea and argue that it is impossible for the researcher to be a “blank slate” (Suddaby, 2006; Timonen et al., 2018). However, the grounded theory researcher should be careful not to bring in any theoretical assumptions and force the data into them, they should be transparent and open to unexpected findings. One of the elements of traditional research approaches that might introduce

theoretical assumptions to a GTM study is the literature review (Andrew, 2006). Nonetheless, a literature review was necessary for this study in order to put the study into context. To limit the risks presented by this challenge, I split the literature review process into a foundational literature review, done at the beginning of the data collection, and integrated one done after and during data analysis. Furthermore, the foundational study only focused on general conceptualizations, and a broad field of study, without going deep into one theoretical realm. Lastly, at the beginning of this chapter, I stated my position as a researcher in this study and tried to be as open and transparent as possible.

Finally, CGT is a relatively new methodology, without many use-cases, and homogeneity. Some CGT scholars base their methodology on the classical grounded theory approach (Looker et al. 2021), and others on the constructivist one (Hadley, 2019). Navigating through such a heterogeneous field proved to be challenging for a researcher who is not experienced in using GTM. It was of crucial importance to avoid methodological confusion, especially relating to the data analysis and coding processes. In order to avoid this, I had to choose one approach, and come back to it for further guidance. For this study, the coding and data analysis process is based on the work of Hadley (2019).

Chapter 4: Methods - Sampling, Data Collection and Analysis

In this chapter, I will discuss the methods of the study or the design of this research. In particular, I will present the sampling strategy, data collection methods and data analysis methods. In a grounded theory study such as this one, these processes are highly interconnected and all happen simultaneously. However, for the sake of clarity, I have decided to separate the three into different sections. The order of the separate sections does not imply that they happened chronologically in that order.

Three levels of sampling

The sampling for this research was done in three stages, two stages on the sampling of data sources, and an additional stage on selecting the two companies to be investigated as cases.

Selecting the two cases: Blackboard and Coursera

It is important to note that even though this is a study of two cases, it does not attempt to fully apply the methodological principles of case study research. The methodology of this study is solely led by CGT, which in turn, is compatible with case study research. Fernández (2004), further argues that the combination of these two is not only compatible but rewarding.

The case selection process for this thesis was informed by the relevance to the research questions, my own experience in the field, the foundational literature review, and previous academic work on case study research. Additionally, for the purpose of this study, the selected cases had to have enough open and publicly available data regarding their products and business practices. The cases, Coursera and Blackboard, were selected taking three criteria into account: they are somewhat *typical* examples of online education companies, they are *influential* in their field of operation, and they are *different* from one another. Firstly, the two cases are deemed typical because both companies have conventional organizational structures, and both cases have typical business models in the online education industry. Secondly, they are influential because they are one of the 'Goliaths' in their respective industries. Coursera is one of the biggest MOOC providers, and Blackboard is one of the most popular Learning Management System (LMS) providers. Nevertheless, Seawright and Gerring argue that an influential case cannot be considered typical, since if it was typical it would not have commanded all that influence (2008). However, I believe that these cases are typical, not because they are not influential, but rather because they are, and as such, they dictate the mainstream of their respective fields. Lastly, the two cases are different from each other in their product offerings, target customers, and relationship to educational institutions. For instance, whereas Blackboard is largely used in formal education scenarios and is usually employed by large academic institutions such as school systems or universities, Coursera is often used by individual learners in a non-formal educational setting. In order to understand the cases more clearly, I will present short 'case profiles' in the following two sections.

Coursera Inc.

Coursera is one of the biggest and most well-known MOOC providers. Coursera was founded in 2012 by two Stanford University computer science professors Andrew Ng and Daphne Koller. As such, alongside EdX and Udacity, Coursera is considered one of the pioneering MOOC providers. With a revenue of over $293 million U.S. dollars, Coursera is the biggest MOOC provider when it comes to market share (SEC, 2021). Additionally, Coursera is the only MOOC provider that has held an Initial Public Offering (IPO). Coursera partners with educational institutions (e.g. universities) and corporations (e.g. IBM, Google, Microsoft) to create educational content in the form of online courses, and digital degrees. Coursera's business model includes multiple revenue streams:

- *Certificates.* The sale of certificates verifying the student's completion of a course on Coursera is the company's first and most basic revenue stream
- *Specialisations*. Specialisations are a series of courses that in the end form a unit accounting to some professional competency (e.g. graphic design, conflict management). Whereas the content of specialisations is free, the ability to earn grades and receive feedback is paid.
- *Coursera for Business.* Coursera for Business is the enterprise, Business-to-Business (B2B) corporate e-learning product of Coursera.
- *Coursera Plus.* Subscription model for Coursera's Specialisation courses and certificates. Coursera offers an annual or monthly subscription to access unlimited content and earn unlimited certificates as an alternative to paying separately for single learning units.
- *Coursera for Governments and Non-profits.* Coursera's product that is tailored to governments and large non-profit organizations.
- *Online Degrees*. Coursera offers multiple online bachelor's and master's degree programs and qualifications. Online Degrees on Coursera can cost up to $50,000.
- *Professional Certificates*. Partnering with big corporations such as Google and IBM, Coursera also sells professional certificates for around $39 per month. The courses are created by the corporations and are supposed to supply them with a workforce for their specific IT, data science or marketing needs.

Coursera's main business model is attracting learners with free content and selling a premium service once the learners are engaged with that content (Coursera, 2021).

Blackboard Inc.

Blackboard is one of the largest educational technology companies. Blackboard is particularly well known as a provider of learning management systems such as Blackboard Learn. The company's main client base consists of education providers such as universities, large corporations, and governments. A testament to Blackboard's popularity is the 16,000 organizations from over 90 countries that use their products and services. Blackboard's business model is based on selling access to its products and services such as learning management systems, data analysis software, web conferencing and others. Furthermore, Blackboard is known for its aggressive expansion and acquisition strategies. Products and services include:

- *Learning Management*. Blackboard's most popular product is Blackboard Learn, an LMS used by nearly a quarter of U.S. schools. Additionally, other learning management products that Blackboard offers include a mobile learning solution called Blackboard App, online teaching software named Blackboard Instructor, and the popular plagiarism prevention tool, SafeAssign.
- *Data and Analytics*. Blackboard offers multiple data collection and analysis tools and services such as Blackboard Reporting, a tool that tracks learners' behaviour and usage on LMSs, or Blackboard Intelligence a student data management system
- *Blackboard Collaborate*. A virtual classroom web conferencing tool.
- *Recruitment Services*. Blackboard provides academic institutions with digital marketing, enrolment, and research services, so these institutions can increase enrolment and attract learners.
- *Student Success Services*. Blackboard provides student success such as the Retention Support, a data-driven solution that improves learner retention.
- *Consulting Services*. From educational to technical guidance, Blackboard provides multiple consulting services.

Initial and Theoretical Sampling

The sampling for this thesis included two stages, the initial sampling, which was done before the data collection process, and theoretical sampling that was conducted simultaneously with the data collection and analysis.

Prior to any data collection, a combination of initial data sources was identified. The initial data sources were selected based on the relevancy to the initial research question, my judgment informed by previous research experience in the field, and public availability of the data. Because most of the companies dealing with big data keep their practices in a "black box", the open and public access to data was a crucial deciding factor for selecting the specific data sources. Privacy Policy and Terms and Conditions documents were the first identified data source. They are publicly available, partially answer the initial research question, and in my judgment are a good start for diving deeper into the data practices of these companies. Similarly, press releases were considered as well. The short list of initial data sources included Privacy Policy and Terms and Conditions documents, press releases, and Website and Social Media Copy.

The moment I began collecting data from the initial three data sources, the data analysis process began, and with that the second stage of sampling, theoretical sampling. The codes and categories emerging from the initial data sources informed the selection of new data sources. The choice of new data sources was informed based on the existing gaps in knowledge in the data, the need for further explanation, or the emergence of categories that can be supported by an additional data source. In order to more clearly understand the theoretical sampling process, I will present a few examples of how data sources were added to the study.

Throughout the beginning stages of data collection and analysis, from the privacy policies and terms of condition documents, a category pertaining to the sharing of data with third parties emerged. Understanding that Blackboard and Coursera are sharing data with third parties, a further explanation was necessary for understanding how this shared data was used by the third-party partners of

Blackboard and Coursera. Therefore, “Documents by Third Party Partners” was added as a new data source to the scope of the study.

Moreover, to further support the presence of a category named “Use of learner-generated data for business development”, I had to understand what exactly was entailed by business development. Thus, “Budget and Earnings Reports” was added as a new data source.

The rationale behind selecting certain data sources is also recorded in the Memos that were generated throughout the study. For instance, Memo#19 explains the rationale behind using Amazon Web Service’s (AWS) Privacy Policy as a data source.

Memo #19:
5/10/2021: After realizing that Coursera and Blackboard both share student data with third parties, I decided to look more into this. One party, or partner of both companies that struck my eye was Amazon. Particularly, both Coursera and Blackboard are AWS users. Knowing Amazon's 'loose' data practices, I decided to look more into the AWS T&C

Table 1.1 *Memo #19*

The full list of data sources includes:

Data Sources	Source Type 1	Source Type 2	Source Type 3
Contemporary and Historical Documents	Privacy Policy	Terms and Conditions	Cookies Policy
Website and Social Media Content	Contemporary/Historical	App Market Discriptions	User signed in/No user signed in
Secondary Interview and Publically available data from key actors	CEO/CTO/CMO/Data Scientists	Media Interviews	Blog posts and Social Media
Press Releases and Reports	Press Releases	Budget and Earning Reports	Product Update Reports
User Experience Observation	Going through a Course	Engagement emails	Notifications
Third party partners	AWS T&C		

Table 1.2 *Data Sources and Types*

Data Collection Processes

For the purpose of data collection, I used two tables, one for Blackboard Data and one for Coursera data. Each table had five columns, and each row would represent one data point. By the end of the study, there were a total of 158 data points. The first column is *data type*, this would include the data source category and the type. The second column is the *quote*, this includes the exact textual content of the data point, or if the data source is an observation the full explanation of the observation. The third column is the *summary and description*, here I either briefly summarise or describe the content of the quote. I tried to write descriptive summaries, rather than my thoughts and opinions regarding the quote, for this, I used memo writing. The fourth column is reserved for coding and data analysis, such as writing codes, emerging categories or concepts. The fifth and last column is the *source* of the data, meaning where the data was found. The source of the data would most often be a URL address. Table 1.3 presents a snippet of a data collection table that was used.

Data Type	Quote	Summary & Disctiption	Code	Source
Documents; Privacy Notice	Coursera, Inc. is the data controller of the personal information we collect about you (i.e., the entity that determines the means and purposes of collecting, using, and disclosing the personal information), unless you are part of a degree, certain MasterTrack programs, or certain other circumstances as communicated to you, in which case Coursera is the data processor.	Coursera is the entity that decides what is the learners data used for. If users are part of a degree, then the institution is the data controller	**1. Data control agreed between institutions and Coursera**	Key Inform https://ww
Documents; Privacy Notice	We collect the personal information... including account registration details such as name and email, details of Content Offerings you undertake, survey information (where you provide it), identity verification data, and information about your use of our site and Services.	Which personal data is gathered	**1. Unclear wording**	Key Inform https://ww
Documents; Privacy Notice	We use your personal information for the purposes set out... including providing our site and Services to you, ensuring the security and performance of our site, conducting research relating to Content Offerings, **sharing information with our Content Providers and our suppliers, direct marketing, and performing statistical analysis of the use of our site and Services.**	What is the data that is gathered used for	**1. Commodifying Data** **2. Sharing with 3rd Parties**	Key Inform https://ww
Documents; Privacy Notice	In order to access certain features and benefits on our Site, you may need to submit, or we may collect, **"Personal Data"** (i.e., information that can be used to identify you and which may also be referred to as "personally identifiable information" or "personal information"). Personal Data can include information such as your name, email address, IP address, and device identifier, among other things. You are responsible for ensuring the accuracy of the Personal Data you submit to Coursera. Inaccurate information may affect your ability to use the Site, the information you receive when using the Site, and our **ability to contact you. For example, your email address should be kept current because that is one of the primary manners in which we communicate with you.**	Personal data is necessary in order to provide the service.	**1. Focus on Privacy not on economic value of data**	What Infor https://ww

Table 1.3 *Example of the data collection table*

Due to the vast amount of data available, and the fact that all the data used in this study is publicly available on the internet, the process of data collection was more about data selection, rather than collection (Bowen, 2009).

The methodical collection of data from documents such as the company's privacy policy, or cookie policy proved to be fruitful. These data sources accounted for almost half of all the data selected for this study. This was expected because only in documents like these, these companies are legally obliged to disclose their big data practices without having much freedom to align the wording with their marketing or corporate interests. From the Privacy Policy, the Terms and Conditions, and the

Cookie Policies of the two companies, I collected and recorded every section that related to the use of learner's data. Furthermore, as the study progressed and categories started emerging, I went back to these documents and collected data related to these emergent categories. This would include data that either supports or challenges the existence of the emergent category.

Similarly, I also selected data from websites and social media content, press releases, and budget reports. I chose to collect data from these sources because they inform about how these companies communicate their big data practices on one hand, to learners and the public, and on the other to investors and potential business partners.

Conducting observations was the most challenging data collection process. To conduct observations and collect data, I went through a single course on Coursera, collected marketing and product emails that Blackboard and Coursera sent me as a user, and recorded notifications that were shown to me while being on these companies' websites and using their products. However, I failed to collect a significant amount of relevant or useful data due to the fact that big data often is about big populations. Having access to only one small part of the plethora of communications, nudges and content that is shown to all the users of Blackboard and Coursera prevented me from seeing the bigger picture.

Nevertheless, for the purpose of the study, I also selected a number of key actors in order to collect and analyse their statements regarding the use of big data and learning analytics in their respective companies. Statements were collected from media interviews of the key actors, articles and social media content they have written, or talks they have given as part of an event. The key actors were selected based on two criteria. Firstly, their position in the company was taken into account. Persons that were on senior-level positions in Coursera and Blackboard, such as the CEOs, Chief Technology Officers (CTOs), Chief Marketing Officers (CMOs), and other C-level executives were selected. Secondly, persons that worked on developing and designing big data tools and products for these companies. This includes data scientists, team leads of the analytics departments, Data Science VPs and managers etc. Data from these key actors was crucial for understanding the underlying logic behind big data applications and practices.

Data Analysis

The Data Analysis and coding processes in this dissertation are largely informed by the Critical Grounded Theory (CGT) methodology formulated by Hadley (2017), which is further expanded on by the same author in 2019. In the beginning stages of data analysis, this methodology mainly draws on Constructivist grounded theory as seen in Charmaz (2000) and Classical Grounded theory developed by Glaser. However, during the later stages, particularly in theory construction, Hadley adopts aspects of Straussian grounded theory as well (Hadley, 2019). However, even though Hadley's CGT is inspired by previous grounded theory approaches, there are some unique methodological distinctions. Namely, there are four methodological stages in CGT research; Open Exploration, Focused Investigation, Theoretical Construction, and Transformative Dissemination (2017b; 2019).

Open Exploration

The Open Exploration stage is the starting point of the CGT study. It begins with the "abstract wonderment of what is going on" (Glaser, 1992, p.22), and a reflection of what the author already thinks they know about the field of curiosity. This reflexive part was conducted by having an "imaginary interview" with myself, which was then used for comparison with the emergent concepts and categories in the later stages of the study (Hadley, 2019, p.17). After this reflexive stage, the initial data collection and coding begins. The data analysis approach here is almost identical to Constructivist and Classical grounded theory approaches. Firstly, I wrote summaries of my observations and the textual initial data, and simultaneously, I started writing memos and coding. Besides Glaser's standard question of "What is going on here?" (1978, p.57), I also asked the question 'Why is this happening?', in line with Charmaz (2008). Furthermore, the open coding process and the memoing in the open exploration stage was also guided by questions specific to critical grounded theory as laid out by Hadley (2019, p.21). In order to more appropriately fit the area of study, I slightly modified some of the questions proposed by Hadley (2019). Some most notable questions are:

- How are things of value being gained or lost here?
- What is being done to gain dominance over others?

- How is power and justice distributed and used here?
- Who has been excluded from the processes of value creation and distribution here?
- What 'invisible' actions do the 'excluded' perform?
- Who benefits and who loses most from what is going on here?

The initial codes and memos that emerged from the data, guided by these questions facilitated the emergence of new research problems, questions, sampling directions, and hypothetical categories. For instance, the second research question, stated in the introductory chapter, was largely formed in this stage. This allowed for the study to progress to the next stage of data analysis, Focused Investigation.

Focused Investigation

Even though comparison between the codes and categories happens constantly throughout all stages of the study, focused Investigation is the stage where the open codes are purposefully compared to one another in order to group them and identify emergent categories in the data. In this stage, I also started using the scholarly literature as another source of data, and another body that can further inform the focused coding. The emergent categories were either created based on one existing, dominant code that had other codes supporting it, or new categories were created that bound multiple codes together. The stage of focused investigation was also guided by the previously posed questions. For instance, from one open code, of an interview of the Chief Content Officer of Coursera, who spoke about using behavioural sciences to drive "learner success", the category "Behavioral Monitoring and Engineering" emerged. On the other hand, the category "Vendor-Institutional Complex", encompassed multiple focused codes relating to the economic relationship between educational technology vendors (Coursera and Blackboard) and academic institutions (schools and universities).

One of the most defining factors of the Focused Investigation stage was the constant and frequent use of theoretical sampling. As categories emerged and connections between the codes were drawn, the need for new data and new sampling directions materialised in order to either support or challenge these categories.

As a product of the data analysis processes in the Focused Investigation stage, a total of eleven categories appeared. However, by the end of the study, these categories were delimited and only three categories remained standing. These categories formed the basis for the continuation of the study to the next stage, Theoretical Construction.

Theoretical Construction

Theoretical Construction is the most challenging and abstract part of CGT. It is the penultimate stage of the study and the culmination of the grounded theory building endeavour. During theoretical coding, I connected the three categories and defined the relationship between them, which allowed for the emergence of one core category, the basis of the emergent theory and my thesis. To support the choice of these three categories, and to support the rationale behind the relationships and connections, further, final theoretical sampling and data collection was conducted. During the process of theoretical construction, I extensively used diagramming and visualisation of the connections between the categories. This helped me both understand, and communicate the categories and the emerging theory better, as shown in the next chapter.

Furthermore, even though retroduction was used throughout the whole study, it was most evident and utilized during this stage. Retroductive inquiry allowed me to explore "what must be taking place in order for the theoretical coding and dimensions surrounding the concept to be true." (Hadley, 2019, p.23). By means of retroduction, I came to ask myself, 'what else is there that sustains this emergent theory?'. From this retroductive process, the last research question stated in the introductory chapter was formed. This allowed me to critically consider the social causes, processes, and contexts surrounding the core category. From this retroductive inquiry, another category emerged, the "Magic Trick" that holds the theory together. At first, I struggled to place this category in the conventional methodological matrix, thus I named it a sustaining category. The sustaining category is not particularly connected to the core category or to any of the other categories of the emergent theories, but it is the invisible surface that supports the theory as a whole. In more abstract terms, it is the glass table where all the categories and connections between them are laid out. In order to attempt and seek

solutions to the problems and issues present in the developed theory, one must carefully inspect and deconstruct the sustaining category.

Transformative Dissemination

The Transformative Dissemination stage of the CGT study is where I am finding myself now. It is the stage of writing up and communicating the theory. To successfully communicate the theory, one must place it into context and integrate it with contemporary scholarly literature. Furthermore, the grounded theorist, in this case, I, should be transparent and explain my background and my position in this research (Hadley, 2019). I did this at the beginning of the previous chapter. Furthermore, as part of the Transformative Dissemination stage, I also evaluated the emergent theory and discussed its implications. The evaluation and discussion of the theory are presented in Chapters 5 and 6. Lastly, the critical grounded theorist should not only comment, explain and uncover oppressive social processes, to be truly critical in a constructive manner one must also propose possible solutions and modes of resistance (Hadley, 2019).

Chapter 5: Research Findings and Theory Building

This chapter presents the findings of the thesis and the emergent, substantive theory that was developed during the data analysis and coding process. The research findings bring together results from the data analysis, relevant data codes, memos, and emerging categories in order to construct a conceptual framework of 'what is going on' in the field of big data in online education, specifically relating to the critical issues regarding economic fairness and digital labour. The core category of *Exploitation of the learning community*, the constituent concepts such as; *the Vendor-Institutional Complex, Use of learner generated value for profit,* and *the Behavioral monitoring and engineering;* and the sustaining category, *the Magic Trick*, were the foundational findings that will serve as the base for the construction and presentation of the substantive theory.

The theory is presented by using multiple visualisations and diagrams in order to more clearly communicate and contextualise the concepts and the relationships between them. As the theoretical concepts and the relationships between them are presented, the scholarly literature is concurrently

integrated in order to contextualise and inform the emergent theory in light of contemporary academic works in related fields. This will lead to the formation of a well-integrated, cohesive theory that will be evaluated at the end of the chapter.

Core Category: Exploitation of the learning community

The core category of *Exploitation of the learning community* materialised as a product of the conceptual relationships drawn between the other three surrounding concepts that emerged from the data. In other words, the Core Category is the aggregation of the three main categories or concepts that emerged from the data. Furthermore, it is the central thesis of this research and the basis for the emergent theory.

To more clearly understand the Core Category, I will divide it into two main constituent sections: Exploitation and Learning Community. Furthermore, I will explain how codes in the data, memos and other concepts support the existence of these sections. Lastly, I will try to concurrently integrate each section with existing scholarly literature

Exploitation

This section of the category addresses the question 'what is being done?', it focuses on the action or the practice of exploitation in online education. Exploitation can be defined as the action of taking an unfair advantage over someone, for one's own benefit (Stanford Encyclopaedia of Philosophy, 2016) . Since in the field of big data in online education, data is primarily used for financial gain, we can classify the exploitative data practices in online education as primarily of a commercial character. Therefore, this form of exploitation entails extracting value from vulnerable or unaware individuals and groups in an unfair way, and using this value to generate profit.

Through their data practices, policies, and actions, both Coursera and Blackboard engage in such extraction of value to secure financial gains. The extraction of the value is mainly done through the use of *learner generated data for profit.* Furthermore, what makes this extraction unfair is the *behavioural monitoring and engineering* that supports this extraction, and the *magic trick* that maintains the exploited in a state of unawareness and confusion. An example of the blend between

using learner generated data for profit, and using behavioural engineering to support extraction of value by Coursera is presented in Code #78.

Data Type	Quote	Source
Key Actor: Emily Sands, VP of Data Science, Coursera 2020 Virtual Conference	For example, our learner-product interest models determine what degrees each user sees in their browser, how degrees are ranked in her megamenu and more. These algorithms are built on a deep understanding of learners from self-reported features like work and education history, to behavioural features like how the learner found Coursera, what she searched for and enrolled in, and how she progressed through her learning experiences. Combined with meta-data on each degree and using as training data the conversion behaviour of the millions of other learners who have been exposed to degrees on Coursera in the past, we estimate each learner's interest in each program. ...This is leading to a 40% increase in degree applications through browse.	Sands, E. (2020, April). Coursera's Product Leadership Presents: Product Innovations [Product Innovation Presentation]. 2020 Coursera Virtual Conference, Mountain View, CA, United States. https://www.youtube.com/watch?v=oVXBL8Zv0uU

Table 2.1 *Coursera Code #78*

As highlighted in the excerpt of Code #78, Coursera uses a combination of data from individual learners, and metadata about millions of other learners, to decide what the learner should see on their browser in order to influence them towards enrolling into, and paying for an online degree offered by Coursera. Therefore, that data in this case is an extracted value that is being used for financial profit. The fact that the learners are unaware that they are being shown specific content that influences them towards enrolling into an online degree and spending money, and furthermore the fact that they cannot opt out of being influenced unless they never use Coursera again, makes the extraction of value from data unfair. Additionally, once the learner enrols in an online degree program, their data is being extracted and used to retain them in the programme, motivate them, and support them in order to continue learning. However, by continuing to be enrolled in these programmes, they are also continuing to pay and produce data. This is pointed out in Memo #14, highlighting the difference in communication with the public and with investors.

Memo #14
Both Coursera and Blackboard use learner generated data in order to retain learners on their platforms. Often, in their website content or by key actors this is referred to as improving learner success, increasing motivation or increasing retention (e.g. Codes #54 & #56). However, when discussing user retention in documents targeted at investors, such as Earning Calls, Coursera communicates retention in terms of profit. By retaining learners, Coursera also secures payments for themselves and their institutional partners, additionally, they can extract more data to help their data-driven marketing practices (e.g. codes #69, #70, #71).

Table 2.2 *Memo#14*

Literature on the exploitation of learners in online education in light of the big data revolution is minimal but still existent. For instance, Marshall (2014), raises similar concerns, arguing that a key consideration when investigating a MOOC is whether that MOOC was built primarily for commercial, personal, or institutional goals, or to truly educate learners. Marshall further argues that offering a for-profit educational platform, with financial growth as the primary driver of the business model can lead to unethical decision making (2014). This is specifically relevant to the case of Coursera.

Furthermore, in the digital healthcare field, Lupton (2014), raises similar concerns regarding the exploitation of patients who share their experiences on online health forums. Lupton finds that many of the patients sharing their experiences are not aware that their data is used for commercial purposes, and coins the term ‘the digital patient experience economy’. Similarly, these concerns can be applied to the field of online education. Especially, knowing that learner data from online forums and discussion boards are collected and used for product and business development purposes. This is particularly evident in Blackboard Code #9, and Coursera Codes #14 and #17.

Data Type	Quote	Source
Documents; Privacy Notice	We collect data about your responses to quizzes, your assignments and other course work, and files you submit or upload as well as your activity and actions within our products and services. In some products, you can also provide comments in **discussion forums** and chats and send messages to your peers and instructors. If you are an instructor, we also collect information about your grading, feedback and assessments, and similar actions within our products.	Blackboard; End users: Information we collect https://help.blackboard.com/Privacy_Statement#end-users

Table 2.3 *Blackboard Code #9*

Data Type	Quote	Source
Documents; Privacy Notice	We may share general course data (including quiz or assignment submissions, grades, and **forum discussions**), information about your activity on our Site, and demographic data from surveys operated by us with our Content Providers and other business partners so that our Content Providers and other business partners may use the data for research related to online education.	Coursera; How We Use the Information: https://www.coursera.org/about/privacy

Table 2.4 *Coursera Code #17*

In the wider scholarly work on big data ethics, especially in works in the sphere of economics, exploitation is a commonly discussed term. Zuboff states that the essence of the exploitation in surveillance capitalism is to represent and minimize our experiences to nothing more than behavioural data for the sake of "others' improved control of us." (Zuboff, 2019, p.94). When translated into the field of digital education, the essence of the exploitation is the rendering of learning as behavioural, market, and research data for the sake of increasing the profit of commercial online education providers.

Learning Community

Provided that Exploitation is an existing reality in the field of online education, and big data is the enabler, it is crucial to understand who are the exploited, and why. This section particularly addresses these questions. There are different actors in the big data economy of digital learning. Namely, there are the companies such as Coursera and Blackboard, academic institutions such as universities and

schools, teachers and content providers, and lastly, the learners. In the cases of Blackboard and Coursera, the companies and the academic institutions are the owners and controllers of the data, and they decide how and why the data is used and collected. Blackboard Code #8 and Coursera Code #2, clearly state the power held by Coursera and Blackboard when it comes to data:

Data Type	Quote	Source
Documents; Privacy Notice	We provide most of our products and services to end-users of an institution as a so-called 'data processor' on behalf of our clients (for example, school, districts, universities, and corporations). This means that the main responsibility for data privacy compliance lies with your institution as the 'data controller.' It also means that your institution's privacy statement governs the use of your personal information (instead of ours). Your institution determines what information we collect through our products and services and how it is used, and we process your information according to your institution's instructions and the terms of our contracts with your institution.	Blackboard; End users: Information we collect https://help.blackboard.com/Privacy_Statement#end-users

Table 2.5 *Blackboard Code #8*

Data Type	Quote	Source
Documents; Privacy Notice	Coursera, Inc. is the data controller of the personal information we collect about you (i.e., the entity that determines the means and purposes of collecting, using, and disclosing the personal information), unless you are part of a degree, certain MasterTrack programs, or certain other circumstances as communicated to you, in which case Coursera is the data processor.	Coursera; Key Information: https://www.coursera.org/about/privacy

Table 2.6 *Coursera Code #2*

The data that the companies and institutions control is mined from the learners' activities, content and experiences. Additionally, as mentioned above in Blackboard Code #9, data from instructors and teachers such as their feedback, grading and communications is also being collected. Therefore, we arrive at having two groups with a clear and distinctive difference in power and economic benefit. On one hand, we have the companies and institutions as data controllers who extract value and use it for their own benefit, and on the other, we have the learning community which is comprised of learners and instructors, whose data is being collected.

The learning community, just by existing and functioning on learning platforms such as Coursera and Blackboard, is the producer of big amounts of behavioural and learning data. As producers of such data, the learners and instructors are not compensated for the economic value they are producing, therefore, engaging in invisible unpaid labour. Moreover, a large learning community is both a key selling point for business partnerships and an essential competitive advantage. Therefore, the learning communities are not only the uncompensated producers of data but also the products and commodities of online education platforms. Lastly, due to *behavioural monitoring and engineering*, the learning community are also the subjects in light of big data usage by educational platforms. As such, they are being manipulated, researched about, and experimented on, in order to gain business or product insights, or compel them into paying and producing more data on these platforms. This is represented in Coursera Code #62, where Emily Glassberg, a Data Scientist Manager at Coursera explains the use of data in business decision making research and behavioural engineering.

Data Type	Quote	Source
Data from key actors; Emily Glassberg Sands, Data Science Manager at Coursera, Interview	...The first is our decision science work – developing and testing hypotheses that are key to our product and business direction. Data helps us make decisions that lead to a better experience for learners – whether we're deciding which course topics to source for our catalogue, how to prioritize outreach to learners across markets and languages, or what product changes we can make to help learners stay motivated through the learning journey.	InsideBigData. 2017. Interview with Emily Glassberg: https://insidebigdata.com/2017/02/28/interview-emily-glassberg-sands-data-science-manager-coursera/

Table 2.6 *Coursera Code #62*

The broader scholarly critical literature on the topic is largely compatible with the presented findings in this section. For instance, the learning community is closely related to the 'generators' in Zwitter's taxonomy of big data stakeholders (2014). Moreover, Williamson (2017), adopts a similar position to what is presented above as well. Whereas the findings of this paper state that the learning community is being engaged in free, invisible labour, there is no consensus on this in the literature. Multiple critical authors argue that the data utilizers and collectors benefit from the free labour provided by the producers of data, and oppose this practice (Ritzer & Jurgenson, 2010; Scholz, 2012; Terranova, 2000). Nonetheless, Srnicek (2017a), argues that data is a raw material extracted from the experiences

and behaviour of users, however, the labour is done by the data scientists who employ analytical processes to turn this raw material of data into a valuable, meaningful product. Similar views are expressed by Zhaojun Zhang, a Senior Engineer at Coursera (Code #43), stating that "Data is only valuable when it provides business value.", and business value is derived from accurate data analysis (2016). However, the positions against the existence of 'free labour' fail to recognise the unprecedented case of big data, where the 'raw material' is produced by other human beings, rather than independently existing in nature.

Concept 1: Use of Learner Generated Value for Profit

The *Use of Learner Generated Value for Profit* is one of the central, and first categories that emerged from the data. In its essence, it is the idea that online education providers such as Coursera and Blackboard use the data produced by the learning community for their own commercial benefit. This benefit can be segregated into three goals: *Marketing and Business Development, Research and Partnerships,* and *Product Development*. As shown later, in Blackboard Code #57, these benefits can often synergize and come together, increasing the profit for the education providers.

Marketing and Business Development

Blackboard and Coursera, as the controllers of data and online education providers, are able to translate the learner-generated data into profit by extracting valuable insights that fuel their business development and marketing strategies, or in the case of Blackboard, the marketing strategies of their partner institutions. For Coursera, this can range from internal marketing efforts, such as converting non-paying learners on their platform into paying customers for a low cost of acquisition, to external behavioural advertising methods in order to attract more learners to their platform. Code #78, presented in the previous section of this chapter is a clear example of the former. Furthermore, Coursera Code #75, explains how data-driven algorithms and recommendations are used to choose the best paid degrees to show to learners that come in for a free course. These algorithms have a great effect on lowering the cost of acquisition of paid learners.

Data Type	Quote	Source
Data from key actors; Jeff Maggioncalda, CEO at Coursera, 2020 Virtual Conference	'"Once the learners come in for a free course, our data science team has figured out how to use algorithms and data driven recommendations to figure out, for this learner, this might be the right degree for them, and they've gotten very, very good at it. So, in 2019, the average cost of acquisition for a (we have about 10k learners in our partners' programs right now)... the average cost of acquisition for traditional online program managers is probably $20000 per student, on Coursera the average cost of acquisition is 1250$..."	Maggioncalda, J. (2020, April).Coursera Keynote from Jeff Maggioncalda [Keynote Speech]. 2020 Coursera Virtual Conference, Mountain View, CA, United States. https://www.youtube.com/watch?v=hSn8pe_Cai8

Table 2.7 *Coursera Code #75*

The use of learner generated data to improve Coursera's marketing strategy is evident in their Cookie Policy. Cookies are small text files that contain data that can be verified and traced by web servers (Kaspersky, 2021). They are usually used to identify individual users and track their browsing activities and visits to particular sites. Cookies, for instance, can track whether a user accessed a website by clicking on an advertisement shown on Facebook or Twitter, and how much time they have spent on the website. Moreover, cookies can track whether users that visited the website through a Twitter ad are likely to convert to paying users. Using cookies and gathering data from its 82 million users, all of which are tracked, Coursera has the ability to sharpen their marketing strategy to the most minuscule detail. For instance, they might infer that learners who come through Twitter ads, spend at least three minutes on the website, and visit the About Us page at least once are most likely to pay for a degree. Having this in mind, through behavioural engineering and visual content selection Coursera's marketing strategist and data scientists will do everything that they can to firstly make the user click on a Twitter ad, then visit the About Us page, and spend at least three minutes on the website. The use of Cookies by Coursera for advertising purposes is evident in Coursera Codes #32 and #33.

Data Type	Quote	Source
Documents; Cookies Policy	In addition to our own cookies, we may also use various third-party cookies to report usage statistics of the Site and refine marketing efforts. - Tracking cookies. Follow on-site behaviour and tie it to other metrics allowing better understanding of usage habits. - Optimization cookies. Allow real-time tracking of user conversion from different marketing channels to evaluate their effectiveness. - Partner cookies. Provide marketing conversion metrics to our partners so they can optimize their paid marketing efforts.	Cookies Policy; Third Party cookies: https://www.coursera.org/about/cookies

Table 2.8 *Coursera Code #33*

Additionally to using user-generated value for marketing purposes, Coursera also uses learner data for the development of its business and exploring new profit-making avenues. For instance, as stated in Coursera Code #48, through learner data powered decision making Coursera informs its business development roadmap.

Data Type	Quote	Source
Data from key actors; Vinod Bakthavachalam, Data Scientist at Coursera Website Content; Blog	At Coursera we use data to power strategic decision making, leveraging a variety of causal inference techniques to inform our product and business roadmaps	Bakthavachalam, V. (2018, November). Controlled Regression: Quantifying the Impact of Course Quality on Learner Retention. Medium. https://medium.com/coursera-engineering/controlled-regression-quantifying-the-impact-of-course-quality-on-learner-retention-31f956bd592a

Table 2.9 *Coursera Code #48*

Similarly to Coursera, Blackboard also uses learner-generated data for marketing purposes and behavioral targeting and advertising. Blackboard Code #17 is a statement of this in Blackboard's Privacy Notice.

Data Type	Quote	Source
Documents; Privacy Notice	Marketing. When you use our trial versions, we may use your information for marketing purposes. If you use products and services that we provide directly to you as an end user, such as Blackboard Assist for end users, we may use your information for behavioral targeting of advertising, as stated in the Marketing Section.	Direct individuals; How we use? https://help.blackboard.com/Privacy_Statement#trial-users

Table 2.10 *Blackboard Code #17*

Furthermore, as stated in Code #59, Blackboard also provides digital marketing data-powered services to academic institutions. The key selling point for this service is the ability to closely track learner behaviour through the enrolment marketing funnel.

Data Type	Quote	Source
Website Content; Blog	Tracking and optimizing – Digital marketing allows institutions to track prospective-student behaviour in real-time and see results throughout the entire marketing and enrolment funnel. When configured correctly, these insights enable higher ed marketers to accurately measure and optimize their strategies in order to achieve the highest yield possible—something traditional advertising methods like radio, billboards, and television have long struggled with.	https://blog.blackboard.com/navigating-shifts-online-higher-ed-marketing/

Table 2.11 *Blackboard Code #59*

Research and Partnerships

Besides marketing and business development, the learner-generated value in the form of data is also being used to conduct experiments and research in the newly established field of online education. For this purpose, learners may be shown different variations of content offerings in their courses. Coursera Code #17 explains how this research is often coupled with building profitable relationships with academic institutions and other business partners.

Data Type	Quote	Source
Documents; Privacy Notice	Research. We may share general course data (including quiz or assignment submissions, grades, and forum discussions), information about your activity on our Site, and demographic data from surveys operated by us with our Content Providers and other business partners so that our Content Providers and other business partners may use the data for research related to online education.	"How We Use the Information": https://www.coursera.org/about/privacy

Table 2.12 *Coursera Code #17*

Blackboard shares research data with partner institutions similarly to Coursera. However, Blackboard also uses this learning analytics research for product innovation and development. Thus, synthesizing the general research in online education, and particular research that mostly benefits Blackboard for their product promotion and development. This synthesis is specifically expressed in Blackboard Codes #18 and #57.

Data Type	Quote	Source
Documents; Privacy Notice	We may disclose aggregate or de-identified information that is no longer associated with an identifiable individual for research or to enhance and promote our products and services. For example, we may share aggregated or de-identified information with our partners or others for business or research purposes like partnering with a research firm or academics to explore how our products are being used and how such data can be used to enhance our functionalities and further help our clients and other educational institutions. We will implement appropriate safeguards before sharing information, which may include removing or hashing direct identifiers	Vendors, Partners and Other types: https://help.blackboard.com/Privacy_Statement

Table 2.13 *Blackboard Code #18*

Adding to Code #18, Code #57 gives a depiction of the amount of data and innovative technologies that are used for conducting research. Furthermore, it gives an example of how research and product development are not only compatible but also profitable for Blackboard and their institutional partners.

Data Type	Quote	Source
Website Content; Blog Key actors; Timothy Harfield; Senior Product Marketing Manager for Analytics at Blackboard Inc.	Research is important. At Blackboard, we are in a unique position to be able to conduct learning analytics research at tremendous scale; using cutting-edge techniques on very large data sets that span a broad variety of institutional types. As an educational technology company, we are also able to put our findings into practice through market-leading product innovation. Today, we are thrilled to announce the addition of "course archetypes" to our flagship learning analytics product, Analytics for Learn. This announcement is the most recent example of how research and product development can come together to create something really special.	Harfield, T. (2017). Analytics for Learn: Using Data Science to Drive Innovation in Higher Education. Blackboard Blog. https://blog.blackboard.com/analytics-for-learn-data-science-innovation-higher-education/

Table 2.14 *Blackboard Code #57*

Product Development

There are multiple examples of Coursera and Blackboard using user-generated data to fuel their product development and improve their products. For instance, Coursera has developed a relevancy-based algorithm for their search engine using data of over 10 million learners. This algorithm allows Coursera to show the courses and degrees that learners are most likely to enrol in and pay for. Emily Sands, VP of Data Science at Coursera explains this in Code #80.

Data Type	Quote	Source
Key Actor: Emily Sands, VP of Data Science, Coursera 2020 Virtual Conference	'We also evolved our search engine... to a relevance-based algorithm. Ranking according to what learners searching for that term ultimately went on to enrol in, pay for, and apply to. This enables learners to find the right content from among the vase selection on Coursera faster powered by the search and downstream behaviour of the 10s of millions who came before.	Sands, E. (2020, April). Coursera's Product Leadership Presents: Product Innovations [Product Innovation Presentation]. 2020 Coursera Virtual Conference, Mountain View, CA, United States. https://www.youtube.com/watch?v=oVXBL8ZV0uU

Table 2.15 *Coursera Code #80*

As previously mentioned, there's a certain level of synergy between the different uses of learner generated data for profit. Blackboard Code #57, is one example of this. However, as seen in the Coursera Codes #48 and #80, product development, and marketing and business development are tightly connected to one another as well.

With the three constituent parts explained, we can move onto shortly summarising this Concept and integrating it with the existing literature. The main notion of this Concept is that value is being generated by learners in the form of data, which is then used by Coursera and Blackboard for their own profit and benefit. Furthermore, even though these companies continue to reap the financial rewards of the value generated by learners, the learners are not compensated.

This issue is largely overlooked by the scholarly work on big data in online education, nevertheless, several authors in the relevant academic literature raise similar concerns. For instance, Shum and Luckin (2019) argue that tracking and quantifying human behavioural data is a gold mine for marketers and researchers, but little is being done to improve teaching and learning. Furthermore, Williamson (2019) conceptualises the marketisation of Higher Education and the data infrastructure that surrounds it. Drawing on Srnicek (2017), Williamson brings to light the generation of value and profit from learner produced data (2019). Williamson expands on this by examining the market-making practices in digital platforms in Higher Education, particularly the case of Pearson (2021). Lastly, relating to the *Research and Partnerships* segment of this concept, Marshall (2014), brings up concerns regarding the experimentation on learners using untested pedagogical practices on the EdX online education platform.

Concept 2: Behavioral monitoring and Engineering

If the first concept discussed in this thesis addressed the unfair use of data in online education, the concept of *Behavioral Monitoring and Engineering* pertains to the unfair extraction of data. Behavioural data is the data gathered by tracking and monitoring the actions and experiences of learners, such as how long do learners spend on certain pages, where do they click, what actions do they perform before paying for a course, once enrolled in the course, what steps do they take before

dropping out or successfully finishing etc. Therefore, behavioural data is of central value for online education providers. Blackboard and Coursera use behavioural data for two distinct purposes. Firstly, behavioural data is used to predict and improve what is deemed to be learner success, and secondly for commercial purposes, such as influencing a learner to pay for a certificate or enrol in a degree.

The Behavioral Monitoring and Engineering process is split into three steps. First, it starts by collecting the behavioural data. Secondly, predictive analytics are used to predict future behaviour, such as the likelihood of dropping out, or not finishing a course. Lastly, it ends by intervening in order to alter unwanted future behaviour for the benefit of the company, institution, or the learner. For instance, as seen in Blackboard Code #17, Blackboard collects and uses behavioural data for advertising and marketing purposes. Additionally, a more learner-centred application of this process is most simply and clearly presented in Coursera Code #47.

Data Type	Quote	Source
Key Actor: Helen Zou, Senior Product Manager at Coursera	Because Coursera's courses are often populated by thousands of students, it's not always possible for instructors to keep tabs on each individual student to ensure they're paying attention and aren't falling behind. That's where Coursera uses data from Pendo and other sources to monitor usage and engagement then nudge students onto the paths that the most successful students follow. For example, the data might show that students who watch the first video lecture within five minutes of starting the course are significantly more likely to complete it. Coursera would then send a message to students after enrolment to direct them to that first video, Zou said.	Zou, H. (2020, August). How Coursera is scaling its EdTech platform as learning shifts online. (2021, February 10). Pendo Blog. https://www.pendo.io/pendo-blog/how-coursera-is-scaling-its-edtech-platform-as-learning-shifts-online/

Table 2.16 *Coursera Code #47*

Predictive analytics is the practice of using large historic behavioural data sets to train algorithmic models which then predict the behaviour of learners. Put simply, in order to make predictions about a current, individual learner, these models reflect on how similar learners with similar past experiences behaved. For instance, Coursera may use an algorithmic model to predict whether a learner is likely to pay for a certificate at the end of the course based on their performance and behavioural data and the performance and behavioural data of millions of other past learners in that course.

Besides predicting human behaviour, Coursera and Blackboard use big data to alter it by the use of targeted communication and nudges, visual modification and recommendation models, and advertisements. Targeted communications and nudges are automated messages and notifications that aim at intervening in and altering human behaviour. This method has the simplest underlying model of behavioural control, since it largely relies on verbal or textual communication. However, the structure behind when, where and how are these messages and notifications sent, is incredibly complex and based on large amounts of data and computational analytics. Some messages, such as the one mentioned in Coursera Code #67 aim at altering behaviour in order to improve learner success and the learning experience.

Data Type	Quote	Source
Key Actor: Marianne Sobra, Data Scientist at Coursera Website Content; Blog	The model is a neural network that takes as input a wide range of features, including the following: - The learner's past click-through rates for various messages - Her demographics (e.g., gender, age, country, employment level, education level) - Her on-platform behavioural data (e.g., whether the enrolment is paid, browser language, number of completed courses) - Course-level characteristics (e.g., domain, difficulty, rating) Using these features, the model predicts how likely a specific learner is to find a specific type of pop-up message helpful at a particular point in her learning. If it predicts that the message will have a sufficiently positive impact, it triggers the message; otherwise, it holds the message back.	Sorba, M. (2018, August). Deep learning to intervene where it counts - Coursera Engineering. Medium. https://medium.com/coursera-engineering/using-deep-learning-to-intervene-where-it-counts-aab76c7ce8dc

Table 2.17 *Coursera Code #67*

However, others, such as the notification presented in Figure 1 aim at compelling students towards enrolling and paying for online degrees, certified programs or similar paid content.

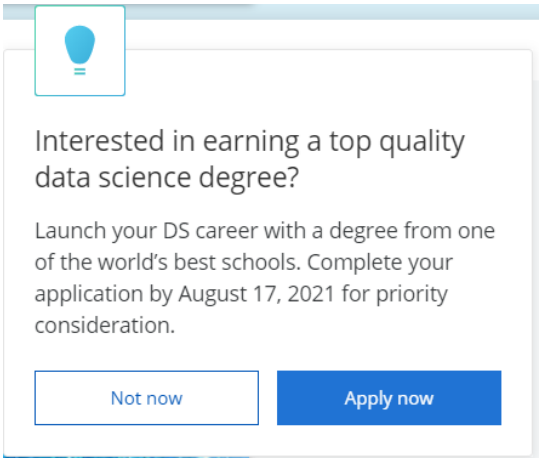

Figure 1 *Coursera Code #81 - Observation of an automated notification*

Furthermore, another mode of behavioural engineering is the visual modification and recommendation models. Often, this mode is also named 'Personalisation of Content'. I will not use this terminology, since I believe it falsely represents the practice of modifying content for commercial benefit as an attempt for personalisation and improvement of the learner's personal learning journey. Through content modifications and recommendations informed by big data, companies such as Coursera can control what the learners see, and do not see. Consequently, learners might enrol in a degree that is just simply made more visible to them, rather than taking their own, personal learning path. This is depicted in Coursera Code #75 where Coursera succeeded in lowering their cost of customer acquisition down to $1250 by using algorithms that control what degrees are recommended and marketed to learners.

Many works in the contemporary scholarly literature deal with behavioural data in online education (Kizilcec et al., 2020; Qiu et al., 2016; Tseng et al., 2016; Wassan, 2015). However, critical perspectives on the use of behavioural data in the field are rare (Regan & Jesse, 2018; Reidenberg & Schaub 2018). Firstly, Reidenberg & Schaub (2018) raise concerns over the increase in learner stress, knowing that their steps are being watched and surveilled. Moreover, they further note the danger of the use of learner behavioural data for manipulation outside of the learning context, for commercial purposes. Similar to the findings in this thesis, Regan and Jesse (2018) find the ethical issues of nudging problematic in certain circumstances, especially in the field of education. They argue that

these nudges must be transparent and promote social welfare, rather than become tools of manipulation for commercial benefit. The findings of this study are further supported by Yeung (2017), arguing that by using nudges companies become "choice architects" and have the power to alter human behaviour in a predictable way.

Moreover, ethical concerns regarding the use of behavioural data are being raised in the broader literature on big data, as well, especially in the fields of information systems and economics (Herschel & Miori, 2017; Newell & Marabelli, 2015; Zuboff, 2015; Zuboff, 2019). For instance, Newell & Marabelli (2015), uncover the falsely portrayed 'free access' to information on the internet, arguing that in fact, large tech companies have control over what we see and access. They further argue that this control over what the user sees leads to a slow and subtle manipulation of the user's worldview.

Concept 3: The Vendor-Institutional Complex

The title of the *Vendor-Institutional Complex* concept is partially inspired by the existence of other industrial complexes such as the Military-Industrial complex, or the Prison-Industrial Complex. It captures how institutions, in this case, academic ones, reconstruct their relationship with industrial enterprises in accordance with capitalist and neoliberal models with the aim of financial growth.

In online education, institutions and vendors (such as Blackboard and Coursera), as owners and controllers of the data, have a shared, vested economic interest in extracting data from learners and benefiting from the free labour that the producers of data provide. Therefore, their relationship forms an economic model that is based on and aimed towards the extraction of value from the data students and teachers produce. Even though both Blackboard and Coursera are engaged in the Vendor-Institutional Complex by partnering with universities and other academic institutions, there is much richer data explaining Blackboard's involvement in such relationships with their institutional partners. For instance, Blackboard Code #52 indicates the underlying logic behind the Vendor-Institutional Complex.

Data Type	Quote	Source
Key Actor: Timothy Harfield; Senior Product Marketing Manager for Analytics at Blackboard Inc.; Website Content; Blog	I believe that it is important because it has the potential to reduce barriers to cooperation between institutions, catalyse community, generate innovative solutions to specific problems, and shine a light on high impact scalable practices that would otherwise not be promoted by institutions themselves. It is the job of institutions to extract value from their data through action. Through Blackboard's dual-investment in products and people, we are helping colleges and universities to do just that	Harfield, T. (2017b). Extracting Value From Data. Blackboard Blog. https://blog.blackboard.com/extracting-value-from-data/

Table 2.18 *Blackboard Code #52*

In order to more clearly understand the Vendor-Industrial Complex, *Figure 2* presents a diagram that I created during the Theoretical Construction stage. I will further explain this diagram by listing the four main steps in the cyclical process of the Vendor-Institutional Complex.

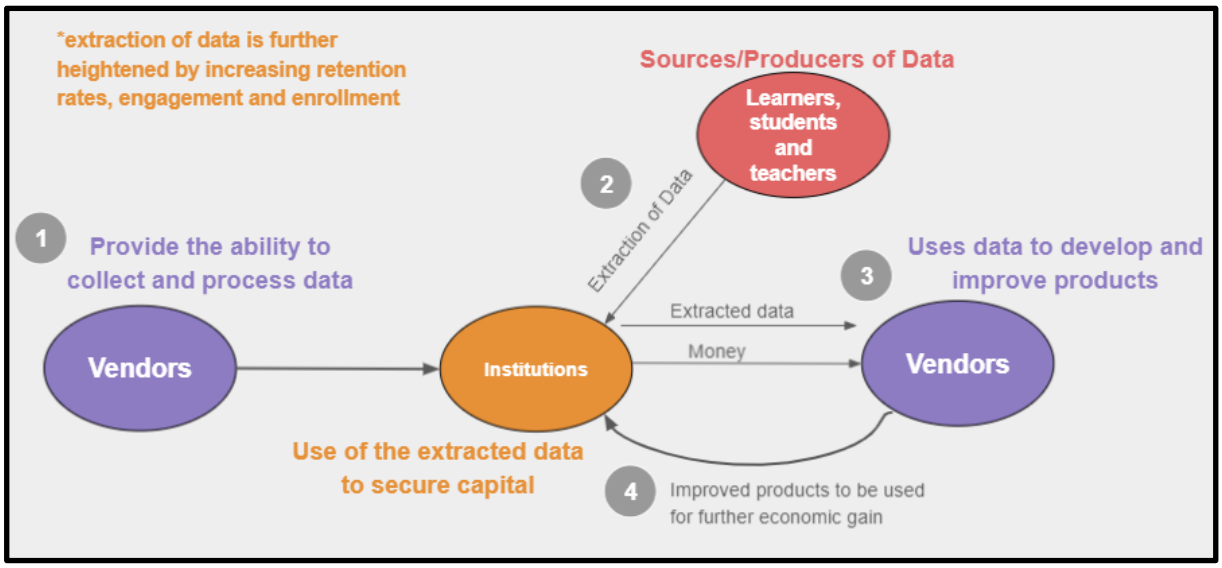

Figure 2. *The Vendor-Institutional Complex*

As seen in the diagram above, the Vendor-Institutional Complex has four main stages or steps.

1. Firstly, the vendors, in this case, Blackboard, provide the ability for institutions to collect and process data *en masse*. This practice is called the productization of data collection and processing. Blackboard Code #49 specifically relates to this step.

Data Type	Quote	Source
Key Actor: Mike Sharkey; Vice President of Analytics at Blackboard Inc.; Website Content; Blog	The purpose of analytics tools is to provide access to data that institutions already have without also limiting access through proprietary algorithms and data models. It also means working with clients to prepare them for long-term success with our data offerings. For us, extracting value from data involves more than unearthing data as a resource. It involves working with our customers to understand their problems, people, and processes, adapting our solution to their unique contexts, and connecting them to our large community of supportive users.	Sharkey, M. (2018, October). Mining Educational Data & Creating Value in Higher Education. Blackboard Blog. https://blog.blackboard.com/mining-educational-data-creating-value-higher-education/

Table 2.19 *Blackboard Code #49*

2. Academic Institutions extract data from students and teachers, who are seen as the mere producers or sources of data. Once the data is extracted, institutions use this data to gain value and secure economic gains. The data can be used for commercial purposes such as cutting costs, retaining students, or improving administrative efficiency, or informing digital marketing strategies. Coursera Code #75, is an example of how data can drive the cost of customer (student) acquisition for an online degree from $20,000 to $1,250.
3. Following the extraction of value, institutions share the extracted data with the vendors and provide them with payment for their services. In turn, Vendors use this data and resources to further develop and improve their products. Codes in the *Product Development* section of the first concept discussed in this chapter depicts how learner data is used to improve and develop new products.
4. Lastly, these improved or newly developed products and services are sold to academic institutions, which are then used for further data extraction, economic gain, and cost-cutting. Blackboard Code #64 explains how innovative insights from learner data are provided to and used by academic institutions.

Data Type	Quote	Source
Key Actor: Kathy Vieira; Chief Portfolio Officer at Blackboard Inc.; Website Content; Blog	We've used these insights to inform innovations that enable personalized experiences on your campus based on aggregate data from around the globe. Through Blackboard Achieve we put these insights into your hands with a unique view into your institution's data so that you can provide your students with a more personalized experience by knowing what's driving student performance in every department, program and course.	Vieira, K. (2020, November). Introducing Blackboard Achieve – the next step for Blackboard Data. Blackboard Blog. https://blog.blackboard.com/blackboard-achieve/

Table 2.20 *Blackboard Code #64*

Lastly, it is important to note the cyclical and reproductive nature of the Vendor-Industrial Complex. The increased efficiency of data practices and improved retention and enrolment rates lead to further data extraction from a larger pool of learners and teachers, or in other words producers of data.

The Vendor-Institutional Complex is a novel conceptualisation and to my best knowledge, does not relate to any of the previous literature. For instance, Reyes (2015), completely excludes vendors and online education platforms as stakeholders that benefit from big data in online education. Furthermore, Selwyn (2014) provides a critical perspective of the 'digital university', arguing the emphasis on neoliberal logic by educational key actors such as policymakers and influencers. However, Selwyn does not explore the role of vendors and private companies in the process of building the digital university (2014). Therefore, critical scholarly work focusing on the relationship between academic institutions and commercial vendors is quite limited, and further work exploring the vendor-institutional complex is needed.

Sustaining Category: The Magic Trick

The Magic Trick category emerged by asking myself "what must be true for the exploitation of the learning community to be taking place, and how is this model maintained?". One of the reasons for this inquiry was because I was confused by the fact that individuals and groups within the learning community are not massively protesting this exploitation. This confusion is evident in Memo #26.

Memo #26
So far, I have constructed a theory of exploitation of the learning community and I have conceptualized how academic institutions and commercial vendors (Blackboard and Coursera) are benefitting from this model. However, I am confused as to where the data producers (learners and teachers) are situated in this context. Particularly, if they are exploited, why is there no major pushback against this model in the learning community. Answering this requires further data collection and analysis specifically pertaining to the perspective of students and teachers in the model.

Table 2.21 *Memo#26*

Upon further data collection and analysis, I arrived at two emerging possibilities. One, the learning community is comfortable with and consensual to the big data practices and the logic underlying them. Two, there is a lack of informed knowledge about the exploitative practices, and these are being hidden from their awareness. The first possibility has some minimal supporting evidence, such as Blackboard Codes #37 and #38, which suggest that students were comfortable with being contacted based on the use of learning analytics. However, the students were not informed about what information was collected and how they were tracked, and the research was conducted by Blackboard. Furthermore, supporting evidence for the second possibility is overwhelmingly more voluminous.

The name of the concept, *Magic Trick*, comes from the three different methods used to conceal the exploitative practices of Blackboard and Coursera. In other words, the learning community is tricked into unawareness. Any good magician uses three basic methods to pull off a magic trick; confusion, distraction, and deception. Similarly, these practices are also present in the *Magic Trick* that big data based online education vendors are playing on the learning community.

Confusion

When magicians perform a trick, they might employ a tactic of overwhelming the subjects with too much information or simply performing a plethora of movements and actions so that the subject is left confused. Confusing and overwhelming the audience is one way of covering what the magician is really doing.

Coursera and Blackboard, virtually employ that same tactic of confusion, by presenting the audience with overwhelming amounts of information that is often unclear, and that is incredibly difficult to navigate. For instance, Memo #29, presents an observation made about the time and effort it takes to go through all the information needed for one to understand how their data is used. Namely, one user needs to go over approximately 100 pages of highly technical text.

Memo #29
Through my data collection and analysis work, I have come to realize how much time and effort is actually needed to clearly understand how Blackboard and Coursera are using learner-generated data. For instance, in the case of Blackboard, one must go through over 50,000 words of text (privacy statements, terms and conditions of use, third party statements etc.), and that is not including the privacy policies and statements of the academic institutions and some smaller third-party partners, who also use user-generated data on Blackboard.

Table 2.22 *Memo#29*

Moreover, it's not only that the amount of text is overwhelming, but the wording in the privacy statements and documents is often incomplete and unclear, leaving open possibilities for further exploitation. This unclear wording, when communicating the collection of data from learners often includes phrases such as "among other things" or, "any other data that is generated by you", setting no boundaries to what data can be collected and for what purposes. These particular cases are evident in Blackboard Code #22, and Coursera Code #8.

Additionally to the overwhelming and unclear information provided to learners by Coursera and Blackboard, learners must also go through the data policies of third-party partners and policies based on local laws and regulations. For instance, one of Blackboard and Coursera's largest partners is Amazon Web Services (AWS), they use learner-generated data on Blackboard and Coursera to train their machine learning algorithms (e.g. algorithms for natural language processing, facial recognition etc.). Moreover, as stated in AWS Code #3 section (a, ii) Amazon reserves the right to move learner data from regions with strict data regulations to regions where they can exploit this data freely. Thus, the data of a European student under GDPR can be stored, processed and used in a region that is not under GDPR jurisdiction, where the student privacy and rights are not protected.

Data Type	Quote	Source
Documents; Privacy Notice	(a) You agree and instruct that: (i) we may use and store Amazon Connect ML Content to maintain and provide Amazon Connect ML Services (including development and improvement of Amazon Connect ML Services) and to develop and improve AWS and affiliate machine-learning and artificial intelligence technologies; and (ii) solely in connection with the development and improvement described in clause (i), we may store your Amazon Connect ML Content in AWS regions outside the AWS regions where you are using Amazon Connect ML Services. By contacting AWS Support and following the process provided to you, you may instruct AWS not to use or store Amazon Connect ML Content to develop and improve Amazon Connect ML Services or technologies of AWS and affiliates.	https://aws.amazon.com/service-terms/

Table 2.23 *AWS Code #3*

Distraction

Often, a magician will want to shift their subject's focus away from what is really important, the trick. They do this by distracting the audience by presenting a dummy point of attention, or a decoy. Unlike economic fairness, safeguarding privacy does not challenge the logic behind the commercial value generation from big data in online education. After all, the financial gains in online education are not made by monetising personally identifiable information, but by productizing and marketizing big data sets and data analysis tools. Therefore, shifting the focus of ethical concern away from the economic exploitation in the field is achieved by paying and driving special attention to privacy. In this study, I have recorded twelve codes where privacy concerns have been addressed by Coursera and Blackboard, however, none addressing concerns over economic fairness and data exploitation. This overwhelming focus on privacy is also translated in the academic literature, where most of the work on big data ethics in online education is focused on privacy issues (Chen & Liu, 2015; Fischer et al., 2020; Johnson, 2014; Prinsloo & Slade, 2017; Reidenberg & Schaub, 2018; Wang, 2016; Williamson, 2017b).

Deception

The last and the most central step of any magic trick is *Deception.* It is the act of leading someone to accept a false truth, or in other words, the act of hiding the truth under a veil of falsehood. Relating to this phenomenon, a peculiar category emerged from the data I collected on Blackboard and Coursera; the synthesis of learner success and commercial gain. Namely, both companies marry learner success with their financial success and the financial success of their partners. This way, Blackboard and Coursera can exploit learner's data by falsely claiming that it is the learner's success that they have in mind, not profit. The synthesis is best presented in Memo #14, which is listed at the beginning of this chapter. Furthermore, Coursera Code #71, most clearly portrays how the false veil of pursuing learner success is used to cover the truthful priority and goal of these companies, financial gain.

Data Type	Quote	Source
Documents; Earnings Call Q1; Jeff Maggioncalda, CEO at Coursera	"Our number one goal has been and always will be to serve learners. Our mission is to provide universal access to world-class learning so that anyone, anywhere has the power to transform their lives through learning. Today I am pleased to report we are delivering on that mission. We grew revenue 64% to 88.4 million dollars. Each of our segments saw strong double-digit trends..."	https://event.on24.com/eventRegistration/EventLobbyServlet?target=reg20.jsp&referrer=&eventid=3111489&sessionid=1&key=6B34976168A5CA3C36D489FE7799099A&regTag=&V2=false&sourcepage=register

Table 2.24 *Coursera Code #71*

In this code, Jeff Maggioncalda, the CEO of Coursera explains to investors that the mission of Coursera is to provide universal access to word-class learning, however, the evaluation of this goal is presented in financial terms of revenue growth, not the number of learners or successful course completions, metrics that would more likely reflect learner success.

An additional category that might be relevant to the phenomenon of *Deception* is the devaluation of data. By arguing that data must first be analysed, refined and cleaned before it is valuable, key actors at Blackboard and Coursera are assigning no value to the raw data that is generated by the learning community. This way, the extraction of learner data will not be perceived as economically unfair, or as exploitation, since these companies are not extracting anything of direct commercial value.

Blackboard Code #47 compares raw data to unrefined oil or gold, stating that raw data must be cleaned and analysed in order to be of any value. However, much like unrefined oil and gold, raw data is still a valuable resource that defines both economic and power relationships. Furthermore, unlike unrefined oil and gold, data is produced by humans, that have agency, rather than bio-chemical processes.

Data Type	Quote	Source
Key actors; Mike Sharkey; Vice President of Analytics at Blackboard	Almost all assets have no direct value. Oil and gold need to be refined. Fields need to be cultivated. Factories need to be staffed. When it comes to data, we need to be prepared to do some heavy lifting before it becomes valuable as something that can help us to achieve a particular goal.	Sharkey, M. (2018, October). Mining Educational Data & Creating Value in Higher Education. Blackboard Blog. https://blog.blackboard.com/mining-educational-data-creating-value-higher-education/

Table 2.25 *Blackboard Code #47*

These three methods of Deception, Confusion, and Distraction converge to form the sustaining category of *the Magic Trick.* Which in turn sustains the current economic model of exploitation by "playing a trick" on the learning community, leaving it deceived, confused and distracted.

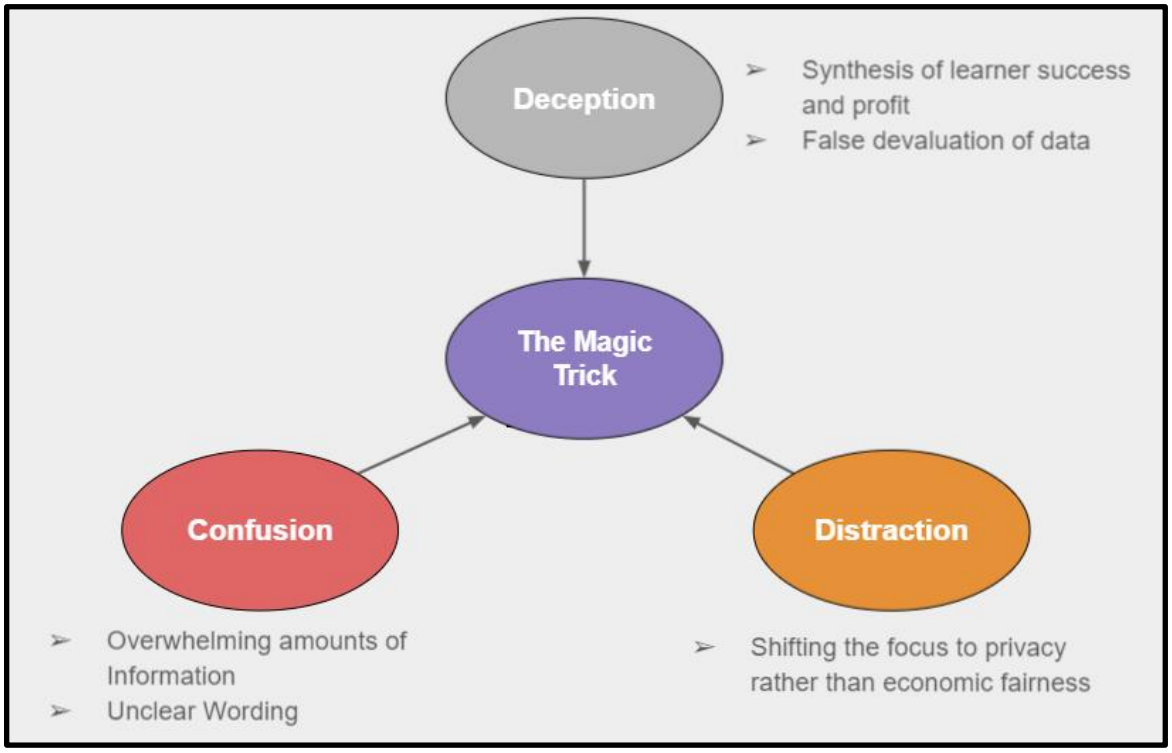

Figure 3 *Sustaining Category: The Magic Trick*

Conceptualizing the Relationships and Summarizing the Emergent Theory

The findings in this chapter presented the emergence of a core category, three main concepts and one sustaining category. These main elements and the relationships between them compose the Theory of Exploitation of the Online Learning Community in the era of Big Data. The theory explains the purpose and role of Big Data in creating and maintaining the economic model and labour relationships in the field of online education. Being critical in nature, the theory particularly raises concerns regarding economic fairness and labour exploitation. Furthermore, by incorporating the sustaining category of *the Magic Trick*, the theory further explains how the economic model and exploitation are maintained. Figure 4 illustrates an overview of the theory. The sustaining category of the *Magic Trick* is not included in the illustration since *the Magic Trick* is the invisible background on which the relationships between the concepts and the core category play out.

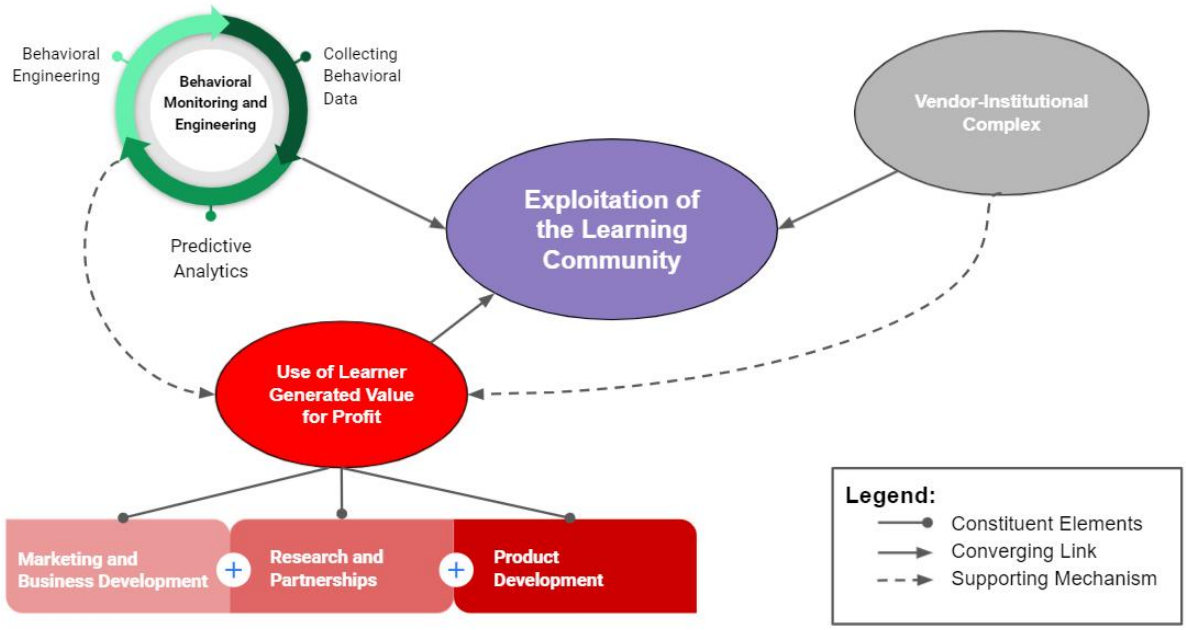

Figure 4 *Theory of Exploitation of the Online Learning Community in the era of Big Data*

By looking at the legend of the Figure, one can notice that there are three types of relationships; constituent elements, converging links, and supporting mechanisms. The first one relates to the elements that constitute a certain concept. For instance, predictive analytics is a constituent element of the concept *Behavioral Monitoring and Engineering.* The converging links represent the convergence of the concepts into the Core Category. In other words, when the main three concepts are united, the core category of *Exploitation of the Learning Community* emerges. Each of the three concepts plays a part in explaining how and why the learning community is unfairly exploited. Lastly, the supportive mechanism links represent relationships where one concept supports the existence of the processes in another. For example, the *Collection and Engineering of Behavioural Data* supports the *Use of Learner Generated Value* for marketing and business development purposes. These constituent elements of Concept 2 provide the necessary mechanisms for the materialisation of the processes in Concept 1.

Evaluation of the Emergent Theory

To ensure that the emergent theory is of good quality and meets the criteria of a solid grounded theory, I have decided to conduct an evaluation of the emergent theory. The process was largely reflexive and personal, rather than quantitative, and following a rigid structure. Different grounded

theory variants have different criteria for assessing the quality of a theory (Charmaz & Thornberg, 2020). Because CGT is a novel branch of GTM and does not have clear evaluative criteria, I decided to use criteria proposed in Classical Grounded Theory work by Glaser (1978). Glaser proposed four evaluative criteria for judging the quality of the emergent theory.

1. *Workability.* The workability of the theory is evaluated through predictions, explanations, and interpretations that it provides regarding the substantive area of study. Since the emergent theory aims at making sense of the complex conceptual and social relationships pertaining to the use of big data in online education, it is an explanatory theory rather than a predictive one. As such, the emergent theory provides viable explanations and interpretations of the social reality in the substantive field.
2. *Relevance.* The relevance of the theory to the people within the substantive field is crucial. A good quality grounded theory should provoke action in the area it explains, by shedding light on hidden problems and processes. This qualitative criterion is crucial for a critical grounded theory. The emergent theory uncovers several generative mechanisms and processes that contribute to the unfair and exploitative treatment of certain groups within the area of big data in online education. Thus, paving the way for actionable change and counter-hegemonic practice.
3. *Fit.* This evaluative criterion pertains to the theory's fit with the data. In other words, a quality grounded theory should fit the "empirical situations" in the substantive area (Lomborg & Kirkevold, 2003, p.191). The presented theory in this study emerged from the data, and its presentation and construction are supported by the relevant codes and categories that were discovered during the data collection and analysis process. Furthermore, I, the researcher, attempted to avoid introducing pre-established theoretical frameworks in order not to force the data in them. Understanding that this is a difficult and nearly impossible task, I tried to be transparent and open about my positionality, previous knowledge and ethical and moral standings.

4. *Modifiability.* As new data emerges, a quality grounded theory should be modifiable and open to changes. Due to the fact that this study employed a limited amount of data sources (e.g. it did not include primary data sources such as interviews), and that the area of study is fast-changing, the possibility that new data will emerge and require modifications to the emergent theory is high. Therefore, even though there is pragmatic value to the theory as it stands, the theory is only provisional, and is flexible to be adjusted and modified as new data and new social realities emerge. Furthermore, the study looked into only two cases as the focus on inquiry, therefore, as the theory is brought into new contexts, it presents possibilities for adaptation and modification.

Chapter 6: Conclusion

Implications

This study presents various social, economic and academic implications. Namely, I identified four possible implications of this study:

1. Uncovering injustice and empowering the exploited. By presenting an explanatory theory that uncovers some hidden and purposefully veiled social and economic realities, the study challenges the exploitative status quo in online education. By uncovering some of these realities, this study aspires to empower the learning community of learners and teachers by providing them with part of the necessary knowledge and tools for anti-hegemonic action. Therefore, this study has strong social implications for activist and social justice movements in the field of online education, digital equity, and data democratisation

2. Promoting social and economic change. By uncovering the realities of the exploited learning community in the field of big data in online education, this study exposes the necessity for change and emancipatory social action. Furthermore, raising issues regarding the fairness of the current economic logic behind the use of big data in online education, the study has strong demands for economic change. Therefore, the study has implications in the movements promoting and fighting for digital economic democracy and equity.

3. Improving conceptual clarity. Despite the existence of multiple works that raise ethical concerns regarding the use of big data in online education, scholarly work in the field lacks a holistic conceptual analysis of the subject. By providing a provisional theoretical framework, and a conceptualisation of the processes and realities related to big data in online education, this study has strong implications for the academic community dealing with these issues. The study invites scholars, students, and the public, to engage with the emergent theory, adapt it to their specific contexts, and further expand it.

4. Methodological implications. Critical Grounded Theory is quite a novel methodology, that has multiple approaches and interpretations. Furthermore, as a novel methodology, CGT lacks the large amount of diverse exemplary work that many other methodologies have. To my knowledge, this is one of the first CGT studies in the field of education that uses only secondary, text-based data sources (most studies use interviews). Therefore, I hope that this study can add to the knowledge and practice of CGT, and encourage other students and learners to use this methodology. Thus, besides the social, economic, and academic implications, this study has strong methodological implications as well.

Limitations of the Theory

Besides the methodological limitations addressed in Chapter 3, there are other, theoretical limitations that demand consideration. Firstly, including only two case studies as the focus for the study, the knowledge and the theory that emerged is local and narrow in context. Therefore, the emergent theory is not, and it does not aim or claim to be generalizable, limiting the applicability of the theory to different contexts. However, as previously mentioned, the theory is modifiable and open for adaptations and comparisons with contexts, different from the one studied.

Secondly, other than my personal experiences and observations, the theory does not include the experiences and knowledge from main actors in the field such as learners, teachers, and employees in online education companies. Dealing with particular themes such as exploitation and deception, these perspectives are crucial for the development of a holistic theory.

Lastly, besides offering a conceptual map that provides opportunities for social change and anti-hegemonic action, the theory and the study itself do not present viable solutions and potential avenues for action. The reason for this is to maintain the scope of the thesis within the recommended institutional guidelines (i.e. word count).

Suggestions for further research

Suggestions for further research arose from the limitations of the study and some conceptual gaps that emerged during the theory-building process. Firstly, as mentioned in Chapter 2, more scholarly work and research is needed on the specific definition, conceptualisation and genealogy of big data in online education.

Secondly, research exploring the perspectives and lived experiences of members of the learning community are crucial for the development of a holistic, explanatory theory. Therefore further research in this direction is necessary.

Thirdly, I suggest further research and investigation into the specific concepts that emerged from the data in this study. For instance, the concept of the vendor-institutional complex in online education is seldom explored and researched. This presents an opportunity for further research and development of the provisional theory presented in this thesis.

Finally, this thesis does not include or propose solutions for anti-hegemonic practice. Work and research on exploring possible alternative economic, social and organisational models for the use of big data in online education is particularly needed.

Concluding remarks

Before finishing this work, I would like to present one last remark regarding the tone and intent of the study. When reading this thesis, and interpreting the emergent theory, one might falsely assume that I am criticizing the use of big data in online education as a whole, or that I am advocating against the use of these technologies. However, what I aim at critiquing and advocating against with this thesis is the underlying, exploitative logic behind the use of big data in online education. Big data, learning analytics and artificial intelligence as technologies have huge potential to be beneficial for both

learners, teachers, and educational institutions. However, for these benefits to materialize, the priority when using them should be the wellbeing and flourishment of learners and improving the learning experience, not commercial goals and financial gains.

References

Amirault, R. J. (2019). The next great educational technology debate: Personal data, its ownership, and privacy. *Quarterly Review of Distance Education*, *20*(2), 55-73.

Andrew, T. (2006). The literature review in grounded theory: A response to McCallin (2003). *The Grounded Theory Review: An International Journal*, *5*(2/3), 29-41.

Bartlett, J. (2018). Big data is watching you | The Spectator. The Spectator. https://www.spectator.co.uk/article/big-data-is-watching-you

Belfrage, C., & Hauf, F. (2017). The gentle art of retroduction: Critical realism, cultural political economy and critical grounded theory. *Organization Studies*, *38*(2), 251-271.

Bhaskar, R. (1978). A Realist Theory of Science Brighton: Harvester Wheatsheaf.

Bhaskar, R. (2002). From Science to Emancipation: Journeys towards Meta-Reality.

Blom, B., & Morén, S. (2011). Analysis of generative mechanisms. *Journal of critical realism*, *10*(1), 60-79.

Bodle, R. (2016). A critical theory of advertising as surveillance: Algorithms, big data, and power. In *Explorations in Critical Studies of Advertising* (pp. 148-162). Routledge.

Bryant, A., & Charmaz, K. (Eds.). (2007). *The Sage handbook of grounded theory*. Sage.

Brown, A., Fleetwood, S., Roberts, M., & Roberts, J. M. (Eds.). (2002). *Critical realism and Marxism*. Psychology Press.

Brown, M. (2011). Learning analytics: The coming third wave. *Educause learning initiative brief*, *1*(4), 1-4.

Burck, C. (2005). Comparing qualitative research methodologies for systemic research: The use of grounded theory, discourse analysis and narrative analysis. *Journal of family therapy*, *27*(3), 237-262.

Charmaz, K. (2000). Constructivist and objectivist grounded theory. *Handbook of qualitative research*, *2*, 509-535.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Charmaz, K., & Thornberg, R. (2020). The pursuit of quality in grounded theory. *Qualitative Research in Psychology*, 1-23.

Chamberlain-Salaun, J., Mills, J., & Usher, K. (2013). Linking Symbolic Interactionism And Grounded Theory Methods In A Research Design. Sage Open, 3(3), 2158244013505757.

Chen, X., & Liu, C. Y. (2015). Big data ethics in education: Connecting practices and ethical awareness. *Journal of Educational Technology Development and Exchange (JETDE)*, *8*(2), 5.

Chiang, R. H., Grover, V., Liang, T. P., & Zhang, D. (2018). Strategic value of big data and business analytics.

Chong, D., & Shi, H. (2015). Big data analytics: a literature review. *Journal of Management Analytics*, *2*(3), 175-201.

Corbin, J. & Strauss, A. (2008). Basics of qualitative research (3rd ed.): Techniques and procedures for developing grounded theory. *SAGE Publications, Inc.,* https://www.doi.org/10.4135/9781452230153

Coursera. (2021). *Coursera, Inc. - Investor Relations*. https://coursera.com. https://investor.coursera.com/overview/default.aspx

Coyne, I., & Cowley, S. (2006). Using grounded theory to research parent participation. *Journal of Research in Nursing*, *11*(6), 501-515.

Cutcliffe, J. R. (2005). Adapt or adopt: Developing and transgressing the methodological boundaries of grounded theory. *Journal of advanced nursing*, *51*(4), 421-428.

Daniel, B. K. (2017). Big data in higher education: The big picture. In *Big data and learning analytics in higher education* (pp. 19-28). Springer, Cham.

Denzin, N. K., & Lincoln, Y. S. (2008). Introduction: The discipline and practice of qualitative research.

Dishon, G. (2017). New data, old tensions: Big data, personalized learning, and the challenges of progressive education. *Theory and Research in Education*, *15*(3), 272-289.

Dobson, P. J. (2002). Critical realism and information systems research: why bother with philosophy. *Information research*, *7*(2), 7-2.

Drigas, A. S., & Leliopoulos, P. (2014). The use of big data in education. *International Journal of Computer Science Issues (IJCSI)*, *11*(5), 58.

Dunne, C. (2011). The place of the literature review in grounded theory research. *International journal of social research methodology*, *14*(2), 111-124.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*(8), 1523-1545.

Elia, G., Solazzo, G., Lorenzo, G., & Passiante, G. (2019). Assessing learners' satisfaction in collaborative online courses through a big data approach. *Computers in Human Behavior*, *92*, 589-599.

Emmanuel, I., & Stanier, C. (2016). Defining big data. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (pp. 1-6).

Fernández, W. D. (2004, July). The grounded theory method and case study data in IS research: issues and design. In *Information Systems Foundations Workshop: Constructing and Criticising* (Vol. 1, No. 22, pp. 43-59).

Finn, M. (2016). Atmospheres of progress in a data-based school. *cultural geographies*, *23*(1), 29-49.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130-160.

Fuchs, C. (2012). Class and Exploitation on the Internet. In *Digital labor* (pp. 219-232). Routledge.

Fuchs, C., & Chandler, D. (2019). *Digital objects, digital subjects: Interdisciplinary perspectives on capitalism, labour and politics in the age of big data*. University of Westminster Press.

Gibson, D. (2017). Big data in higher education: Research methods and analytics supporting the learning journey.

Glaser, B. (1978) T*heoretical Sensitivity: Advances in the methodology of grounded theory*. Mill Valley: Sociology Press.

Glaser, B. G. (1992). *Basics of Grounded Theory Analysis: Emergence Vs. Forcing*. Sociology Press.

Glaser, B. G. (1998). *Doing grounded theory: Issues and discussions* (Vol. 254). Mill Valley, CA: Sociology Press.

Glaser, B. G., & Holton, J. (2004). Remodeling grounded theory. In *Forum qualitative sozialforschung/forum: qualitative social research* (Vol. 5, No. 2).

Glaser, B., & Strauss, A. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Adline De Gruyter.

Glenn, R. A. (2003). *The right to privacy: Rights and liberties under the law*. Abc-Clio.

González, R. J. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. Anthropology Today, 33(3), 9-12.

Goulding, C. (2002). *Grounded theory: A practical guide for management, business and market researchers*. Sage.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*(3), 42-57.

Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (2016). The'big data'revolution in healthcare: Accelerating value and innovation.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191-209.

Hadley, G. (2017). *Grounded theory in applied linguistics research: A practical guide*. Routledge.

Hadley, G. (2019). Critical grounded theory. *The Sage handbook of current developments in grounded theory*, 564-592.

Herschel, R., & Miori, V. M. (2017). Ethics & big data. *Technology in Society*, *49*, 31-36.

Huda, M., Haron, Z., Ripin, M. N., Hehsan, A., & Yaacob, A. B. C. (2017). Exploring innovative learning environment (ILE): big data era. *International Journal of Applied Engineering Research*, *12*(17), 6678-6685.

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). The horizon report 2011. *The New Media Consortium, Austin*.

Johnson, L., Adams, S., Cummins, M., Estrada, V., Freeman, A., & Ludgate, H. (2013). The NMC Horizon Report: 2013 Higher Education Edition. *Austin, TX: The New Media Consortium.*

Johnson, J. A. (2014). The ethics of big data in higher education. *The International Review of Information Ethics*, *21*, 3-10.

Jurkiewicz, C. L. (2018). Big data, big concerns: Ethics in the digital age. *Public Integrity*, *20*(sup1), S46-S59.

Kalota, F. (2015). Applications of big data in education. *International Journal of Educational and Pedagogical Sciences*, *9*(5), 1607-1612.

Kapitanova, K., & Son, S. (2012). Machine learning basics. *Intelligent Sensor Networks*, 3-29.

Kaspersky. (2021). *What are Cookies?* Www.Kaspersky.Com. https://www.kaspersky.com/resource-center/definitions/cookies

Kempster, S., & Parry, K. W. (2011). Grounded theory and leadership research: A critical realist perspective. *The leadership quarterly*, *22*(1), 106-120.

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data & Society*, *2*(2), 2053951715611145.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kitchin, R., & McArdle, G. (2015). The diverse nature of big data. *Social Science Electronic Publishing*, *25*(3), 1-10.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., ... & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900-14905.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, *60*(3), 293-303.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321-332.

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016, April). Big data application in education: dropout prediction in edx MOOCs. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)* (pp. 440-443). IEEE.

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, *24*(3), 149-157.

Lomborg, K., & Kirkevold, M. (2003). Truth and validity in grounded theory–a reconsidered realist interpretation of the criteria: fit, work, relevance and modifiability. *Nursing Philosophy*, *4*(3), 189-200.

Looker, B., Vickers, J., & Kington, A. (2021). The Case for a Critical Realist Grounded Theory Research Design. *DEALING WITH GROUNDED THEORY*, 139.

Lupton, D. (2014). The commodification of patient opinion: the digital patient experience economy in the age of big data. *Sociology of health & illness*, *36*(6), 856-869.

Lynch, C. F. (2017). Who prophets from big data in education? New insights and new challenges. *Theory and Research in Education*, *15*(3), 249-271.

Marciano, A., Nicita, A., & Ramello, G. B. (2020). Big data and big techs: understanding the value of information in platform capitalism. *European Journal of Law and Economics*, *50*(3), 345-358.

Marshall, S. (2014). Exploring the ethical implications of MOOCs. *Distance Education*, *35*(2), 250-262.

Marín-Marín, J. A., López-Belmonte, J., Fernández-Campoy, J. M., & Romero-Rodríguez, J. M. (2019). Big data in education. A bibliometric review. *Social Sciences*, *8*(8), 223.

McGhee, G., Marland, G. R., & Atkinson, J. (2007). Grounded theory research: literature reviewing and reflexivity. *Journal of advanced nursing*, *60*(3), 334-342.

Memmi, D. (2015). Information technology as social phenomenon. *AI & SOCIETY*, *30*(2), 207-214.

Menon, A., Gaglani, S., Haynes, M. R., & Tackett, S. (2017). Using “big data” to guide implementation of a web and mobile adaptive learning platform for medical students. *Medical teacher*, *39*(9), 975-980.

Mitros, P., Paruchuri, V., Rogosic, J., & Huang, D. (2013, June). An integrated framework for the grading of freeform responses. In *The Sixth Conference of MIT's Learning International Networks Consortium*.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datafication'. *The Journal of Strategic Information Systems*, *24*(1), 3-14.

O'Reilly, U. M., & Veeramachaneni, K. (2014). Technology for mining the big data of moocs. *Research & Practice in Assessment*, *9*, 29-37.

Pardos, Z. A. (2017). Big data in education and the models that love them. *Current opinion in behavioral sciences*, *18*, 107-113.

Perrotta, C., Gulson, K. N., Williamson, B., & Witzenberger, K. (2021). Automation, APIs and the distributed labour of platform pedagogies in Google Classroom. *Critical Studies in Education*, *62*(1), 97-113.

Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., & Van Zyl, D. (2015). Big (ger) data as better data in open distance learning. *International Review of Research in Open and Distributed Learning*, *16*(1), 284-306.

Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. In *Big data and learning analytics in higher education* (pp. 109-124). Springer, Cham.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in MOOCs. In *Proceedings of the ninth ACM international conference on web search and data mining* (pp. 93-102).

Raitman, R., Ngo, L., Augar, N., & Zhou, W. (2005). Security in the online e-learning environment. In *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)* (pp. 702-706). IEEE.

Regan, P. M. (2000). *Legislating privacy: Technology, social values, and public policy*. University of North Carolina Press.

Regan, P. M., & Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: Twenty-first century student sorting and tracking. *Ethics and Information Technology*, *21*(3), 167-179.

Reidenberg, J. R., & Schaub, F. (2018). Achieving big data privacy in education. *Theory and Research in Education*, *16*(3), 263-279.

Reyes, J., Morales-Esteban, A., & Martínez-Álvarez, F. (2013). Neural networks to predict earthquakes in Chile. *Applied Soft Computing*, *13*(2), 1314-1328.

Reyes, J. A. (2015). The skinny on big data in education: Learning analytics simplified. *TechTrends*, *59*(2), 75-80.

Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest L. Rev.*, *49*, 393.

Richterich, A. (2018). *The big data agenda: Data ethics and critical data studies* (p. 33-51). University of Westminster Press.

Ritzer, G., & Jurgenson, N. (2010). Production, consumption, prosumption: The nature of capitalism in the age of the digital 'prosumer'. *Journal of consumer culture*, *10*(1), 13-36.

Sbaraini, A., Carter, S. M., Evans, R. W., & Blinkhorn, A. (2011). How to do a grounded theory study: a worked example of a study of dental practices. *BMC medical research methodology*, *11*(1), 1-10.

Scholz, T. (Ed.). (2012). *Digital labor: The internet as playground and factory*. Routledge.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political research quarterly*, *61*(2), 294-308.

Securities And Exchange Commission (2021). *S-1 Registration Statement Coursera, Inc.* https://www.sec.gov/Archives/edgar/data/1651562/000119312521071525/d65490ds1.htm

Shum, S. J. B., & Luckin, R. (2019). Learning analytics and AI: Politics, pedagogy and practices. *British journal of educational technology*, *50*(6), 2785-2793.

Srnicek, N. (2017a). *Platform capitalism*. John Wiley & Sons.

Srnicek, N. (2017b). LSE Lit Fest 2017: platform capitalism by Nick Srnicek. *LSE Review of Books*.

Stanford Encyclopedia of Philosophy. (2016). *Exploitation (Stanford Encyclopedia of Philosophy)*. Https://Plato.Stanford.Edu. https://plato.stanford.edu/entries/exploitation/

Suddaby, R. (2006). From the editors: What grounded theory is not.

Taylor, C., & White, S. (2001). Knowledge, truth and reflexivity: The problem of judgement in social work. *Journal of social work*, *1*(1), 37-59.

Terranova, T. (2000). Free labor: Producing culture for the digital economy. *Social text*, *18*(2), 33-58.

Timonen, V., Foley, G., & Conlon, C. (2018). Challenges when using grounded theory: A pragmatic introduction to doing GT research. *International Journal of Qualitative Methods*, *17*(1), 1609406918758086.

Tseng, S. F., Tsao, Y. W., Yu, L. C., Chan, C. L., & Lai, K. R. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning*, *11*(1), 1-11.

Urquhart, C., & Fernández, W. (2016). Using grounded theory method in information systems: The researcher as blank slate and other myths. In *Enacting Research Methods in Information Systems: Volume 1* (pp. 129-156). Palgrave Macmillan, Cham.

Vaidhyanathan, S. (2009). The Googlization of universities. *The NEA 2009 Almanac of Higher Education*, *72*.

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics [Electronic version]. Educational Technology & Society, 15(3), 133-148.

Wang, Y. (2016). Big opportunities and big concerns of big data in education. *TechTrends*, *60*(4), 381-384.

Wassan, J. T. (2015). Discovering big data modelling for educational world. *Procedia-Social and Behavioral Sciences*, *176*, 642-649.

Wielki, J. (2015). The social and ethical challenges connected with the big data phenomenon. *Polish Journal of Management Studies*, *11*(2), 192-202.

Williamson, B. (2017a). *Big data in education: The digital future of learning, policy and practice*. Sage.

Williamson, B. (2017b). Who owns educational theory? Big data, algorithms and the expert power of education data science. *E-learning and Digital Media*, *14*(3), 105-122.

Williamson, B. (2019). Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education. *British Journal of Educational Technology*, *50*(6), 2794-2809.

Williamson, B. (2021). Making markets through digital platforms: Pearson, edu-business, and the (e) valuation of higher education. *Critical Studies in Education*, *62*(1), 50-66.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118-136.

Zhang, Z. (2016). React-based user behavioral tracking - Zhaojun Zhang. Medium. https://medium.com/@zhaojunzhang/react-based-user-behavioral-tracking-d0b8eaa92671

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, *30*(1), 75-89.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st ed.) [E-book]. PublicAffairs.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, *1*(2), 2053951714559253.