HAW HAMBURG

BACHELOR THESIS

Data ethics in the context of data literacy – an analysis of educational approaches for higher education

submitted in August 2020 by

Phuong Nguyen Hoang Nam

Examiner 1: Prof. Christine Gläser
Examiner 2: Tom Alby

HAMBURG UNIVERSITY OF APPLIED SCIENCES
Department Information
Study program Media and Information

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG

Hamburg University of Applied Sciences

Data ethics in the context of data literacy – an analysis of educational approaches for higher education

Bachelor thesis submitted by

Phuong Nguyen Hoang Nam

Abstract

The worldwide technological revolution with mass adoption of automation and AI has been transforming the role of data in modern-day life and end-users find themselves riddled by the dilemma between convenience and privacy. In this context, data literacy proves to be important more than ever. As literacy paves the way to human knowledge, data literacy opens the door to a world with human-beneficial technology. This research utilized the grounded theory approach to perform an inductive analysis on different materials on data ethics of varying lengths and formats. It aims to point out one or more common themes in existing educational approaches on data ethics in the context of data literacy training. The analysis of nine sources resulted in the identification of a central theme of training on data ethics which is stakeholders' ethical role in education, society, and technology. Stakeholders in education comprise roles such as educators, researchers, students. Stakeholders in society are every individual, including people with high social status such as policymakers, judges, and politicians whose actions have a direct impact on humankind. Stakeholders in technology are not necessarily individuals such as users or developers; they could also be professional societies, philanthropic organizations, and corporations. The findings show that current educational approaches for data ethics in higher education emphasize the awareness of learners and their active role in shaping the world regardless of their disciplines. In the long run, the findings of this research could serve as a starting point for further scientific research on the teaching of data ethics. Teaching approaches need not be stakeholder-centric; however, as proven during the analysis, it is notable to consider the stakeholder's roles and responsibilities.

Table of Contents

List of Figures

List of Tables

List of Used Abbreviations

AAC&U	American Association of Colleges and Universities
ACRL	Association of College and Research Libraries
AI	Artificial Intelligence
CC	Creative Commons
CC BY	Creative Commons Attribution
CC BY-SA	Creative Commons Attribution-ShareAlike
CC0	Public Domain
CHE	Centrum für Hochschulentwicklung
CHT	Center for Humane Technology
CIA	Central Intelligence Agency
DataONE	Data Observation Network for Earth
DEK	Deutscher Ethikrat
DPO	Data Protection Officer
EAD	Ethically Aligned Design
ERC	European Research Council
FLI	Future of Life Institute
GDPR	General Data Protection Regulation
HFD	Hochschulforum Digitalisierung
HKUST	Hong Kong University of Science and Technology
HOOC	Hamburg Open Online University
HRK	Hochschulrektorenkonferenz
ICT	Information and Communication Technology
IEEE	Institute for Electrical and Electronics Engineers
ILCSJ	Information Literacy Competency Standards for Journalism Students and Professionals
ILSHE	Information Literacy Standards for Higher Education
ILSSET	Information Literacy Standards for Science and Engineering/Technology
IT	Information Technology
MIT	Massachusetts Institute of Technology
NSA	National Security Agency
OER	Open Educational Resources
PDPA	Personal Data Protection Act
PDPC	Personal Data Protection Commission
PII	Personally Identifiable Information

RGS	Reform Government Surveillance Coalition
SCC	Standard Contractual Clause
STEM	Science, Technology, Engineering, and Mathematics
T&C	Terms and Conditions
TS/SCI	Top Secret/Sensitive Compartmented Information
UKE	University Medical Center Hamburg-Eppendorf
UN	United Nations
UX/UI	User Experience/User Interface
VPN	Virtual Private Networks

1 Introduction

1.1 Topic relevance

The world is in the midst of the Fourth Industrial Revolution (Industry 4.0) with the mass adoption of automation and artificial intelligence (AI). This revolution introduced the dominant presence of data and its indisputable role in both private and professional life. Humans can now not only be regarded as their own entity, but also as data entities. This transition proposes a problem, as many people are not aware that they and the data they are creating are victims of exploitation. Modern-day life has seen a rise in discussions on data breaches, mass surveillance, and AI malfunctions. The public debate comprises of the most different narratives. Whereas a portion of society willingly adopts new technologies such as AI-based voice assistants and even microchip implantations, many skeptics do not possess smartphones nowadays. End-users find themselves riddled by the dilemma between convenience and privacy.

A technology consumer should understand both the benefits and the impending threats that a product or service could bring in an ideal scenario. Ideally, they are aware of the data being collected and generated, and the potential implications they could face. These are all competencies of data literacy. Thus, data literacy can enable anyone to take a stand in a nowadays data-centric world. Whereas the traditional meaning of "literacy" entails being able to read and write, "data literacy is the ability to collect, manage, evaluate, and apply data, in a critical manner" (Ridsdale et al. 2015). As literacy paves the way to human knowledge, data literacy opens the door to a world with human-beneficial technology.

1.2 Subject of the research

Current happenings show that data literacy is gaining importance as an academic competence. Along with the efforts in establishing data literacy as a basic academic competence, it is crucial to consider its ethical aspect. While data literacy promotes the necessary skill to understand and use data in a critical manner, data ethics is a step further that requires this manner to be ethical. A public survey from Hochschulforum Digitalisierung (HFD), a joint initiative of Centrum für Hochschulentwicklung (CHE), Hochschulrektorenkonferenz (HRK) and the Stifterverband, found out that competence on data ethics plays a

significant role in society (Schüller, Busch & Hindinger 2019). To illustrate, data projects that concern sensitive personal data often require the professional support of data protection experts and data ethicists, who are qualified to perform data security measures such as pseudonymization. Similarly, other scientific research on data literacy has repeatedly mentioned data ethics as a crucial aspect of data literacy, mostly as the competence to work with data and apply data (Manduca & Mogk 2002; Maybee & Zilinski 2015; Ridsdale et al. 2015).

Worldwide, there are various researchers and research groups that are actively contributing to exploring the topic of teaching data literacy in higher education in a variety of disciplines, for instance, social sciences, life sciences, and teacher education (Prado & Marzal 2013; Carlson & Bracke 2015; Dunlap & Piro 2016; Gibson & Mourad 2016; Wolff et al. 2016; Ming & Hui 2018). Regarding training on data ethics, even though some researchers have been vocal about the importance of this topic (Floridi & Taddeo 2016; Mittelstadt et al. 2016), there are fewer empirical findings (Manduca & Mogk 2002; Maybee et al. 2015; Ridsdale et al. 2015; Schüller, Busch & Hindinger 2019).

Until now, training data ethics has been scarce and mostly optional, if at all offered for higher education purposes, even though it would benefit anyone regardless of socioeconomic and educational background (Royal Society 2019b). It is probably because the topic has not yet gained importance in the public consciousness and that people with a higher level of education (e.g., data scientists) are the first to be confronted with the topic. The underlying problems could be a lack of scientific findings regarding effective teaching methods and the ambiguity of data ethics as a subject.

1.3 Research objectives

In the transition towards a data-literate and data-ethical world, there is insufficient research on the implications of data ethics and practical educational approaches. This research aims to bridge this research gap by generating new findings on data ethics that would contribute to the education efforts towards widespread data literacy as a whole scheme. As many training materials on data ethics focus on higher education, it is suitable first to expand research in this area and, eventually, to other fields. The main questions that concern this research are: How can data ethics be taught in higher education? What would be the most effective approaches? What does “data ethics” mean for different audiences, especially for

educators and students? Why is it necessary to teach data ethics in higher education? How do teachings in data ethics differ pedagogically, geographically, and in different disciplines? These are a few questions that could aid in the exploration of a lesser-discussed aspect of data literacy. In the quest for possible answers, the research will focus on analyzing training materials to point out one or more common themes in existing educational approaches on data ethics in the context of data literacy training.

1.4 Research method

Due to the scarcity of relevant literature on data ethics training, which infers that training on data ethics has been largely undocumented, an inductive approach would help discover new patterns in training methods and course goals. The discoveries are essential to conceptualize educational approaches for data ethics in higher education (Gabriel 2013). For this reason, this research aims to utilize the grounded theory approach to analyze different materials on data ethics of varying lengths and formats, including university course module manuals, competence frameworks from institutions, and guidelines from professional societies.

First introduced in 1967 by Glaser and Strauss, grounded theory methodology is an inductive approach to research which emphasizes the process of finding new theories and phenomena over testing existing hypotheses. It was initially designed for sociology, as both Barney Glaser and Anselm Strauss were active researchers and educators in this field. Later on, as Strauss continued to apply and teach the grounded theory approach, he and Juliet M. Corbin collaborated to develop the Straussian branch of the original approach. This work intends to follow the Straussian grounded theory approach (Strauss 1987), which demonstrates the joint effort of Strauss and Corbin (1990, 1998) to follow the constructivist interactionist tradition of qualitative research.

The design of grounded theory procedures facilitates the development of concepts that explain the theories of social phenomena under study (Corbin and Strauss 1990). The inductive nature of the grounded theory approach assists in a research attitude that considers the point of view of the subject under study, in this case, of the educators who designed the curricula. By choosing this approach, the research can concentrate on a thorough analysis of training curricula to extract well-integrated concepts that would best explain how data ethics is being taught in

higher education, rather than imposing set criteria on a somewhat less discussed topic.

1.5 Structure of the research

Before diving deeper into data ethics, Chapter 2 will cover the knowledge basis of data literacy. This part dedicates to clearing doubts about the relevance of data literacy in today's world. Section 2.1 will illustrate the global impact of information and data. Additionally, this part will justify the need for data literacy by naming two unprecedented incidents in recent years, i.e., Edward Snowden's revelation of NSA files and the Cambridge Analytica scandal. Both events shed light on the inner working of unethical practices, the complexity of data manipulation, and the vulnerability of modern technology users.

Next, Section 2.2 will introduce the definitions of data literacy. As a new research field, the definitions of data literacy have been evolving, with new findings constantly amending and modifying their forerunners. Various terms such as statistical literacy, information literacy, data information literacy reflect nuances of data literacy (Schield 2004; Stephenson & Caravello 2007; Qin & D'Ignazio 2010). Even though each term indirectly or directly targets data, a distinction between purely statistical competences and other social competencies such as ethical practice is needed.

Following an overview of the definitions, Section 2.3 will shine a light on the current state of data literacy education in higher education. Depending on the definition, which typically directs towards a specific audience (Crusoe 2016), the focus of data literacy education could vary greatly. Among others, both statistical and ethical competencies are considered competences in data literacy. This section will cover current training offers by universities and organizations worldwide. For example, the Open Data movement, Hamburg Open Online University (HOOC), and Open Educational Resources (OER) are frontline initiatives advocating data literacy. Despite such movements still being relatively modest in quantity, their moderate success gives hope for data literacy.

Chapter 3 will move away from data literacy in general and investigate the specific topic of data ethics. However, as data ethics is a new topic as well as a broad term, Section 3.1 will define the scope of the research object. First, there will be a review on moral philosophy to ascertain the application of ethics in technology. Second, the section will sketch out the three conceptual axes of data

ethics: the ethics of data, the ethics of algorithms, and the ethics of practices (Floridi and Taddeo 2016). Third, Section 3.2 will layout the prerequisites for ethical data practice. In particular, this section will take a brief detour to dig deeper into the topic of ethical oversight to prove that it could be a complex undertaking. Then, the section will report current good practices from governments, professional societies, institutions, and companies. Additionally, as promising as the potentials of better data practice, there are outstanding issues needing resolution in order for current progress to move forward.

In Chapter 4, the research will continue with the analysis of data ethics in the context of data literacy for higher education. Section 4.1 will clarify the research focus. When it comes to handling data, there are different expectations according to each user. In the academic field, ethical data practices would mostly be required for research activities. However, the scope is much more significant. Many majors, especially but not exclusively programming-oriented ones, prepare students for the professional world, where the decision-making process should base on ethical grounds (Maybee et al. 2015). Higher education provides the stepping stone to the professional world, including but not limited to professions in data science. From the beginning, assignments and projects sensitize students to handle data, e.g., survey results or gathering information for essays or presentations. During this phase, students form their work ethics, where data ethics also play a role. Next, Section 4.2 will justify why a qualitative method–the grounded theory methodology–is more suitable for gathering new findings on data ethics. Section 4.3 will explain the methodological approach will be explained in greater detail, i.e. the creation of the data base and methodical procedures.

Chapter 5 will present the analysis results and document every step of the analysis. This chapter will give an account of the analysis from the beginning phase, to the building of categories and subcategories, the identification process of relationships between categories, until the emergence of the central category.

Chapter 6 will evaluate the research findings based on the initial the research objectives. Furthermore, it will discuss the analysis process in terms of insights and challenges. The chapter will provide an outlook on the implications of the research findings and suggest the next steps for future research in the field of data ethics training in the context of data literacy.

2 Data literacy

2.1 Context and relevance

2.1.1 Global impact of information and data

With the emergence of the Internet and social media, we are transitioning to a hyper-connected world. Access to information, which was formerly a privilege of scientists and governments, now has become a universal right as listed under Goal 16.10 in the 2030 Agenda for Sustainable Development: “Ensure public access to information and protect fundamental freedoms, in accordance with national legislation and international agreements” (UN General Assembly 2015). One of the two indicators for this goal entails the measurement of countries that adopt and implement constitutional, statutory, and policy guarantees for public access to information. The Secretary-General's report on the progress towards the Sustainable Development Goals confirmed the number of countries with binding laws and policies giving individuals a right to access information held by public authorities at 127 as of 2019, as visualized in Figure 1. In the past ten years, at least 43 countries have reached this goal, 40% of them in Africa (UN Economic and Social Council 2020).

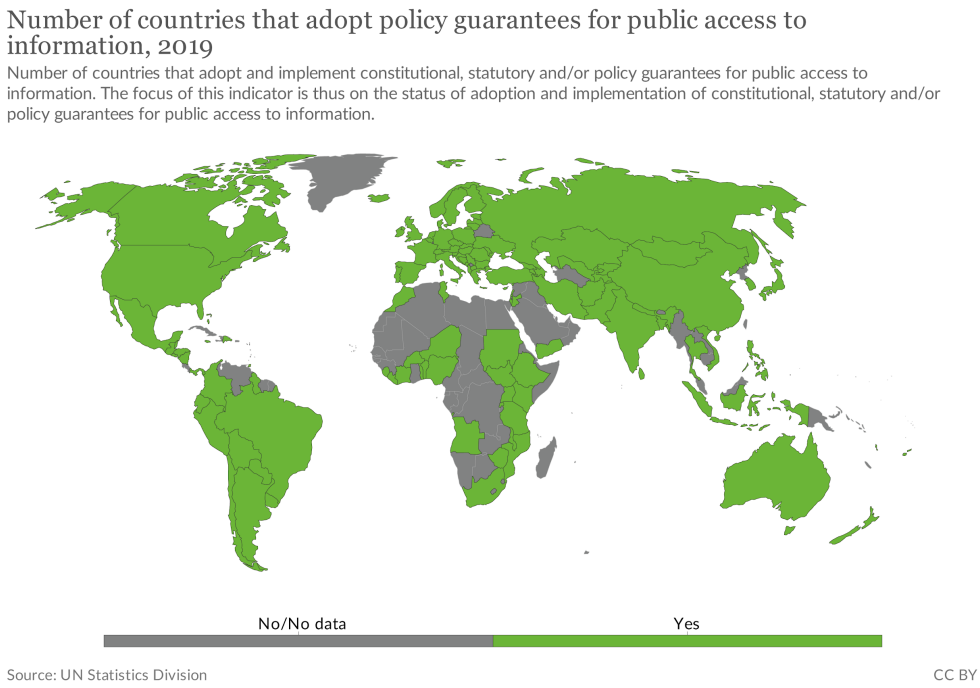

Figure 1: Countries that adopt and implement constitutional, statutory and/or policy guarantees for public access to information (UN Statistics Division 2019)

Before the launch of the World Wide Web in 1992, means of accessing information consisted of mouth-to-mouth knowledge transfer, and then to written records, ranging from engravings on stone caves to books, and along with the emergence of the television, through audio-visual formats. These means have one thing in common: their locality. Even if it is possible to transmit television shows within a vast area, it was unlikely that a person in a particular place could simultaneously watch the same thing as someone else sitting across the globe. The Internet, however, surpassed this limitation. Nowadays, virtually all television shows, radio shows, and other multimedia products could be watched and heard from anywhere in the world, as long as there is a functioning Internet connection. In other words, the Internet has disrupted society and the old way of interpersonal communication (Fuchs 2008).

The "glocalization" of Web2.0 (Boyd 2005) helped diminish the physical and temporal barriers and enabled immediate access to information. While people need a lead time to write a book or produce a movie, the Internet allows instantaneous sharing of information. Some prominent examples are live-streamed webcasts or collaborative tools e.g., Google Docs, which allow multiple users to work on the same document in real-time. For instance, considering academic

information exchange, the Internet has enabled new opportunities (Bik and Goldstein 2013). Besides conferences, congresses, and exhibitions, scientists now can interact with each other on social networks to follow discussions, post content, and discover peers (Noorden 2014). Between face-to-face and online interaction, online discussion facilitates the exchange on a transnational level, regardless of each scientist's geographical location. Needless to say, ease of exchange is not the sole factor in evaluating the effectiveness of online collaboration, nor is it possible to conclude that online collaboration trumps over face-to-face interactions. Other factors must also be taken into consideration, e.g., content quality and user-friendliness. On the other side, in the business world, a survey by IT consulting company Accenture in 2019 revealed that "technologies associated with real-time data capture and analysis" were ranked the most critical technology for transforming/improving business processes (Ghosh, Burden & Wilson 2019).

The discrepancy between two persons regarding access to information lies in the fact that a person may possess all prerequisites for this task, whereas the other does not. Those prerequisites imply that one knows where information is stored and how s/he could retrieve this information. In the context of a hyper-connected world, direct access to an Internet-enabled device, better Internet connection, and a thorough understanding of information-retrieving tools are factors in deciding who can gain better access to information. Among the three, technological competence is critical, as it is possible to purchase better devices and equip faster Internet connection, but data information literacy must be acquired through learning and practice.

As of June 2020, the search term "data" generates about 7,670,000,000 results in the Google search machine. Meanwhile, there are about 154,000,000 results for the search term "data literacy." It is imperative to clarify that data is a broad term with contextual ambivalence. For a computer programmer, data could be binary strings making up programming commands. For a market researcher, data could be demographics collected from a survey. For a politician, data could be polling results. Just as information is vital for every sector and every profession, everyone generates, collects, analyzes, or shares data at one point in their life. Section 2.2 will discuss the relationship between information and data. Indeed, the volume of

data/information created worldwide has been rising exponentially since 2010, as shown in Figure 2:

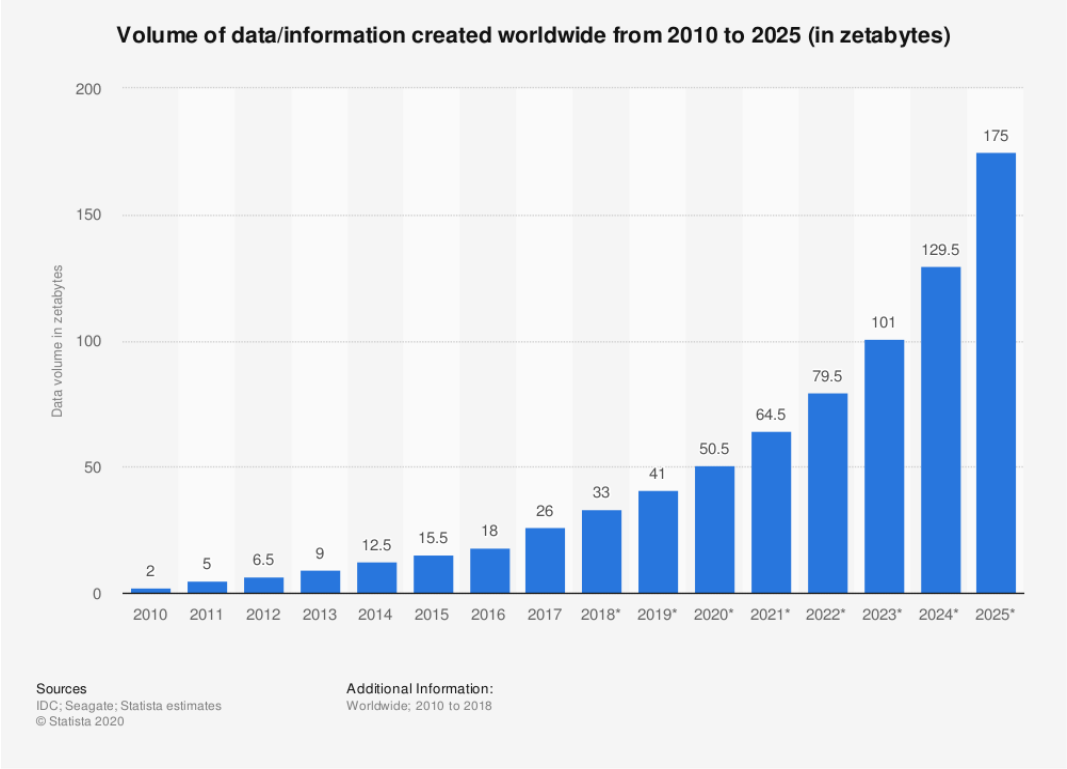

Figure 2: Volume of data/information created worldwide from 2019 to 2025 (in zetabytes) (Holst 2020)

Data has become such a prominent element of modern life that the digital transformation is yielding visible societal changes. The following are a few examples showcasing changes in a data-driven society.

In the business world, companies are spending more and more on digital transformation (Ross et al. 2016). An MIT survey on 179 large publicly traded firms revealed that the output and productivity at companies where data-driven decision-making was adopted are 5 to 6% higher than those of less data-driven counterparts (Brynjolfsson, Hitt & Kim 2011). The consulting industry expected global market volume growth between 2017 and 2020 of 13% to $296 billion, with a high growth rate of 8% in technology consulting, equivalent to $53 billion (Freiland 2016). The position of Chief Data Officers is becoming more popular (Wiseman 2018). A Chief Data Officer makes decisions on the data acquisition of a company. Having been used repeatedly over the last few years, from a cynic point of view, “digital transformation” is a buzzword for a particular management fashion, namely for IT-enabled change initiatives (Abrahamson 1996; Reis et al.

2018). To thrive in the competitive market, businesses and companies will have to continue innovating their operations, adopting new technologies, and investing in human competence (World Economic Forum 2020).

Data, as well as the digital transformation, have such a pronounced effect on humankind, that the emergence of digitization marked the beginning of a new generation called “Digital Natives” (Prensky 2001). Opposed to Digital Natives, Digital Immigrants are born and grew up in a world where digital technology was not substantial, but they are orienting themselves in a digitally-driven world (Palfrey and Gasser, 2008). Nowadays, the emergence and rapid development of social media platforms significantly affect our lives. There are many research findings on the positive and negative impacts of social media platforms on end-users (Kaplan and Haenlein 2010; Ngai, Tao and Moon 2015; Kapoor et al. 2018). There is no need to go into further details to grasp the importance of digital technology in private life.

In the public sector, the digital transformation is bringing in changes to all areas, including government institutions (Janowski 2015), public health management and healthcare services (Gotz and Borland 2016), urban and city planning (Mityagin, Drojjin, and Tikhonova 2017), nonprofit organizations (Beier 2018) and other areas. Not only do databases digitally document patients’ information or insurance information, but they also synchronize this data to ensure efficient patient management. In Germany, in particular, if previously, patients turn to clinics and hospitals when they need diagnoses and treatments, nowadays, health insurance providers such as the Techniker Krankenkasse have a reminder service for its clients to take preventive examinations and request appointments for regular check-ups.

The invention of the Internet and the rapid technological revolution worldwide are bringing drastic changes to modern life. Enabling people with entirely new ways to access information, the digital transformation is happening in every aspect of life, from daily errands to changes in the mentality of a whole generation. Data has become such a prominent element of modern life that the digital transformation is yielding visible societal changes.

2.1.2 The need for data literacy

Besides the positive impacts of the digital transformation on our lives, the risk of unauthorized data sharing has become higher than ever before due to the ease of

data transfer. This risk poses a significant threat to users, as most digital services nowadays rely on the user's data to customize the user experience to leverage their relevance in a highly competitive market. Personally Identifiable Information (PII) is one of the types of information vulnerable to leakage. Krishnamurthy and Wills define PII as "information which can be used to distinguish or trace an individual's identity either alone or when combined with other information that is linkable to a specific individual" (2009). Misuse of PII can lead to devastating implications such as reputation or financial damage.

In June 2013, with the help of journalists from the British news outlet The Guardian, Edward Snowden, a former CIA employee, disclosed approximately 1.7 million documents of secret NSA data on a domestic mass surveillance scheme called PRISM. According to the leak, high-profile companies that collect and store vast amounts of user data like Yahoo, Facebook, Google, Skype, and Apple must oblige to hand over their data to the U.S. government (Greenwald and MacAskill 2013). The following day, he disclosed further information, this time expanding the surveillance scope, particularly an arrangement between the US, the UK, Canada, New Zealand, and Australia. The "Five Eyes," as they were named, allegedly collaborated to monitor transatlantic communications through access to fiber-optic cables, satellites, and radio signals (Beaumont 2013).

On the one hand, this incident is a prime example of what data competence can enable. Edward Snowden explained in his book "Permanent Record" (2019) how he had qualified for the NSA top-secret clearance TS/SCI, cleared the highest clearance in the US called "full scope polygraph," and worked for the NSA at the age of twenty-two without having a bachelor's degree nor an associate's. Thanks to his technical expertise as a system engineer, he had the power to assess some of the most sensitive networks on the planet (Snowden 2019). recognized the problem with the data infrastructure of an institution where data security should have been a priority.

On the other hand, the lack of adequate government reaction to the revelations raised questions from human rights activists about the state of democracy (Jouleva et al. 2013). During a public speech in January 2014, then-incumbent US President Barack Obama defended the NSA program, meanwhile admitted the necessity of specific reforms. Notably, he discredited the influence of Snowden's disclosure (Mason 2014). Following the Snowden disclosure, Rasmussen Reports (2013), Gallup (Newport 2013), and the Pew Research Center (2013) conducted

numerous polls to survey the public's perceptions. There were mixed results, as shown in Figure 3:

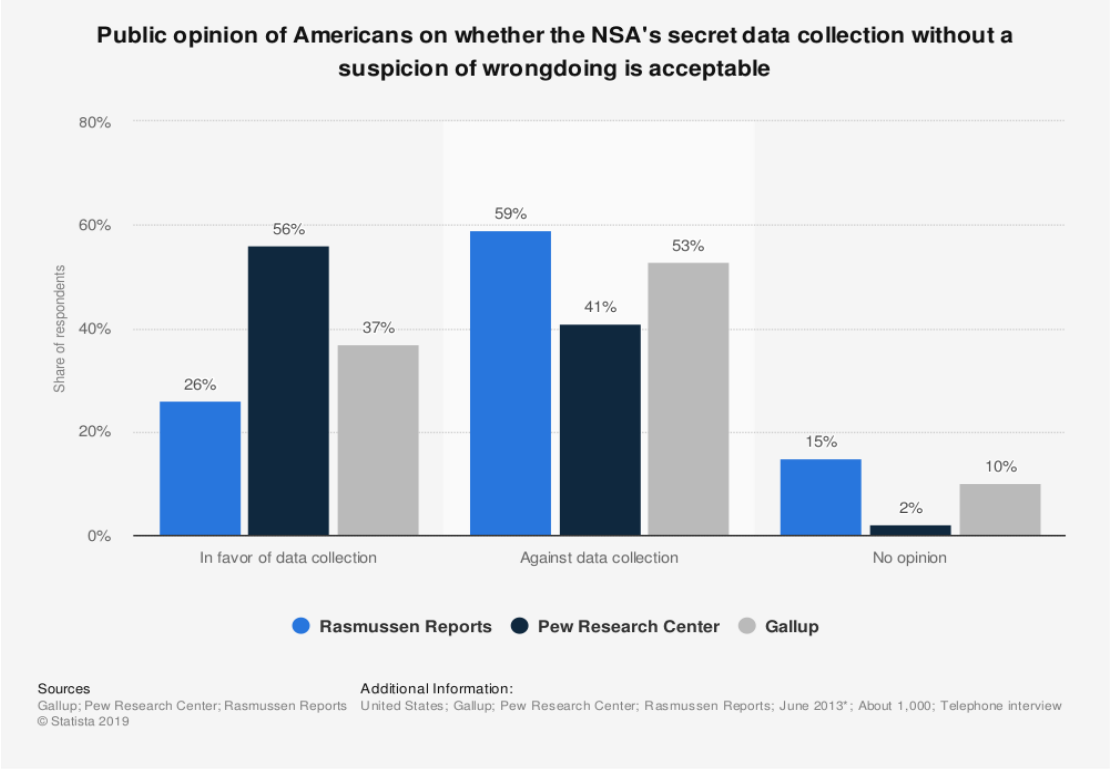

Figure 3: Public opinion of Americans on whether the NSA's secret data collection without a suspicion of wrongdoing is acceptable (Statista 2013)

In 2014, the following year after Snowden's disclosure, a political data analytics firm called Cambridge Analytica gained access to 87 million Facebook profiles by securing a contract with Cambridge psychology professor Dr. Aleksandr Kogan's company (Cadwalladr and Graham-Harrison 2018). Dr. Kogan worked together with the psychologist Michal Kosinski at Cambridge University, where Kosinski developed an online quiz on Facebook to draw users' personality traits based on their activities on the social platform. When a user agreed to take part in the quiz, s/he consented for the app to access their personal information, and also personal data from other people in their social network. After a dispute between the two, when Kogan approached Kosinski with a contract offer from a company called SCL–Strategic Communications Laboratories–to psychometrically gauge millions of American Facebook profiles for undisclosable reasons and faced protest from Kosinski, Kogan went on to register his own company (Grassegger and Krogerus 2016).

It turned out that SCL Group is the mother company of Cambridge Analytica. As early as 2015, The Guardian had already exposed Cambridge Analytica of its activity with Facebook data (Davies 2015). However, Cambridge Analytica first gained significant public spotlight after Donald Trump won the 2016 Presidential Election. In light of this coverage in March 2018, the company faced massive criticism for its data collection, micro-targeting, and manipulating Facebook users for political campaigning purposes (Cadwalladr and Graham-Harrison 2018). In its July interim report, the Digital, Culture, Media and Sport Committee of UK House of Commons made a statement on political advertising via micro-targeting, stating that there should be a ban on micro-targeting political advertising, and a national-level agreement on a minimum limit for individual political messages sent to voters (2018). Being the middleman, Facebook received criticism for having facilitated the data harvesting despite knowing about the potential abuses and received a monetary penalty of £500,000 from the UK Information Commissioner's Office (ICO 2018).

In retrospect, Edward Snowden's disclosure of NSA files (Greenwald, MacAskill & Poitras 2013) and the Cambridge Analytica scandal (Cadwalladr & Graham-Harrison 2018) showed how disrespect to personal data privacy and the unethical exploitation of data could bring about severe irreversible consequences. The incidents shed light on the complicated relationship between tech giants and the governments. Six months after Snowden leaked the NSA files, Google and Facebook, among nine other tech companies, formed the Reform Government Surveillance Coalition (RGS). Its advocacy aimed at reforms that would put the following six principles into actions: 1) limiting governments' authority to collect users' information, 2) oversight and accountability, 3) transparency about government demands, 4) respecting the free flow of information, 5) avoiding conflicts among governments, and 6) ensuring security and privacy through strong encryption (RGS 2020). Six years after the Snowden leak, in 2019, Amnesty International released a report on Google and Facebook, which commented on the RGS and criticized that their "exploitative algorithms" cause grave damages to human rights (Amnesty International 2019).

Even though these incidents unveiled power abuse from influential tech giants and authorities, the weak public response posed a more significant issue. If the general public cannot understand the underlying problem, they cannot comprehend its consequences. The divide in data literacy competences poses a threat to data inequality. While advanced technology like predictive analysis can bring

advantages to a person, it could take away someone else's benefits. The Edward Snowden and NSA files incident explicated that mass data collection for law enforcement purposes could lead to unjust incriminations where people could not object or acknowledge discrimination against them (Katal, Wazid & Goudar 2013). During discussions at a Royal Society and STEM Learning workshop on data science skills in 2019, Dave Gibbs, Senior Computing and Technology Specialist at STEM Learning pointed out that "young people should be encouraged to ask questions about data even if they haven't got the skills needed to process and analyze it. Not everyone needs to become a data scientist, but all young people need the ability to become informed and critical activists, particularly in an era of misinformation and 'fake news'" (Royal Society 2019a).

2.2 Definitions of data literacy

Information is a broad term defined in the Cambridge Advanced Learner's Dictionary & Thesaurus as "facts about a situation, person, event, etc." (Dictionary 2020). The descriptive component making up a piece of information is called data (Griffin 2008). Because of this close relationship, the term "information literacy" often comes up while "data literacy" is discussed. Along with the technological evolution and the ever-evolving understanding of data, the implications behind information literacy and data literacy regularly change accordingly.

Regarding the definition for the term "data literacy", it is essential to consider the relationship between data literacy, information literacy, and statistical literacy. For example, a definition from Schield in 2004 addressed data literacy as an essential component of information literacy and statistical literacy. While information literacy requires the ability to "think critically about concepts, claims and arguments: to read, interpret and evaluate information", statistical literacy requires the ability to "think critically about basic descriptive statistics". Being data-literate means being able to "access, assess, manipulate, summarize, and present data" (Schield 2004, 8). In this context, the object of data literacy is, therefore, quantitative data. Schield is not the only proponent of this definition. Smalheiser published in 2017 an extensive guide for students called "Data Literacy: How to make your experiments robust and reproducible". Having "data literacy" in the title, the book mainly dealt with experiment design and the proper use of statistics. Hence, in this context, data was still understood merely as quantitative data.

Researchers have come up with various definitions for the term “data literacy” (Schield 2004; Stephenson & Caravello 2007; Qin & D’Ignazio 2010). Definitions vary because they depend on the populus, i.e., the relevant population, that data literacy skills can affect. With each different target audience, the proposed definitions emphasize different skill sets. For the same reason, it is necessary to define the populus before making a definition, because it depends on the background and the needs of the populus.

For instance, Crusoe criticized that “the definition for data literacy is myriad and narrow, despite its significance in today’s data-driven world” (2016). He then suggested a definition concerning a large populus as everyone who interacted with or is engaged by data: “Data literacy is the knowledge of what data are, how they are collected, analyzed, visualized and shared, and is the understanding of how data are applied for benefit or detriment, within the cultural context of security and privacy” (2016). This definition echoed that of Prado and Marzal in 2013.

In this context, knowledge and understanding represent data literacy, and the object of data literacy consists of not only numerical data but also non-numerical data that could be transformed into other evaluable forms. Compared to Schield’s definition 12 years earlier, the standards have shifted from statistically evaluating data to understanding the implications of data usage in terms of security and privacy. It is an evolution of data literacy itself that reflects the development of information technology. The more accomplishments humans achieved in information technology, the bigger the need to consider their societal and ethical impacts.

The relationship between data literacy and data-related professions is clear, as one precedes the other. However, data competence should be a fundamental requirement for other professions as well. Anderson, a proponent of “broad data literacy” pointed out that within a corporation, all managers and decision-makers should be data-literate to think in terms of evidence and facts. He stressed that they do not need to possess a deep understanding of data science, rather a general understanding of basic principles to protect themselves from manipulation and statistical biases (Anderson 2015). Anderson’s findings share a middle ground with Crusoe’s definition. Here, “broad data literacy” strengthens the usability of data literacy by illuminating a potential use case in the professional world. This research will lean on Crusoe’s definition of data literacy while keeping in mind the populus of the research scope, i.e., university students.

2.3 Teaching data literacy in higher education

Depending on the definitions that target specific audiences (Crusoe 2016), there are different sets of criteria for competencies. Based on a cluster of interconnected core concepts, researchers can develop competence frameworks. Therefore, a framework suggests flexible options for implementation, rather than forcing a set of standards or learning outcomes, or any prescriptive enumeration of skills (ACRL 2015). The point of competence frameworks for data literacy is to assess understanding and application skills and standardize the evaluation of competences relating to data. Designing a training curriculum with the help of a framework can provide the premise for credible learning outcomes and evaluation.

In the US, the ACRL first introduced Information Literacy Competency Standards for Higher Education (ILCSHE) in 2000. The purpose of this set of standards was to "[provide] a framework for assessing the information literate individual. [The Standards] also [extend] the work of the American Association of School Librarians Task Force on Information Literacy Standards, thereby providing higher education an opportunity to articulate its information literacy competencies with those of K-12 so that a continuum of expectations develops for students at all levels" (ACRL 2000). The American Association of Colleges and Universities (AAC&U), and various discipline-specific organizations have since adopted the standards. In 2012, ACRL declared the need for a revision of the standards, stating that they "do not provide enough guidance on visual literacy and digital literacy, often considered subsets of information literacy itself" (ACRL 2012). The new ILCSHE framework was filed in 2015 and adopted in 2016.

Worldwide, there are various researchers and research groups that are actively contributing to exploring the topic of teaching data literacy in higher education in a variety of disciplines such as social sciences, life sciences, and teacher education. Following is a listing of courses on data literacy in higher education globally, which is in no way exhaustive. Nevertheless, it provides a rough overview of the efforts made by educators and researchers.

In 2008 and 2009, Syracuse University offered a US National Science Foundation course, curriculum, and laboratory improvement grant-supported course in scientific data management (Qin & D'Ignazio 2010). Since 2009, supported by the US National Science Foundation, the Data Observation Network for Earth (DataONE) created educational modules for researchers to "ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-

national science data". It is committed to engaging "students and citizens in science through efforts that span the entire data life cycle, from data gathering, to management, to analysis and publication" (DataONE 2020). In 2012, the Lamar Soutter Library, University of Massachusetts Medical School, and the George C. Gordon Library, Worcester Polytechnic Institute, developed a curriculum framework for undergraduate and graduate students in science, health sciences, and engineering programs. Built upon this framework, the New England Collaborative Data Management Curriculum project was established and led by the Lamar Soutter Library at the University of Massachusetts Medical School (n.d.) in partnership with several libraries in the New England region. In 2014, the University of Minnesota introduced a flipped data management course to graduate students "who seek to prepare themselves as 'data information literate' scientists in the digital research environment" (Johnston & Jeffryes 2014). Stanford University offered a course on using data in journalism (Nguyen 2014). Despite the target audience being students in journalism, the material is also suitable for a broader audience. Carlson and Bracke (2015) performed a case study presenting a student-centered pilot program on data literacy at Purdue University. Through the College of Agriculture, the program was offered and was structured to be flexible enough to incorporate each student's particular field of study.

In the UK, during 2010-2011, the University of Edinburgh, in collaboration with the Institute for Academic Development, ran the project Research Data MANTRA as part of the Managing Research Data program funded by the Joint Information Systems Committee (JISC). This project aimed to reflect best practices in research data management in conjunction with Ph.D. teaching in three disciplinary contexts: social science, clinical psychology, and geoscience (Rice 2011). Later on, the resulting materials were made available through the Institute for Academic Development and EDINA for use by all postgraduate and early career researchers at the University of Edinburgh and made available generally through an open license (EDINA and Data Library, University of Edinburgh 2017). Also funded by JISC in 2011, the DataTrain project started at the University of Cambridge, which aimed to "equip first year postgraduate students with essential skills in looking after their research data for their Ph.D." (Lloyd-Smith 2011) in Archaeology and Social Anthropology.

In 2010, the library at the Hong Kong University of Science and Technology (HKUST) offered a one-credit course on information literacy under the general education free-elective framework (O'Connor & Wong 2010). A team of

librarians taught the courses. The focus was on information literacy; however, there were also data-related topics, such as the use of socioeconomic data. Ting (2015) reported an engineering course of multimedia technology at a vocational institute in Taiwan. Students learned the theories of algorithms and related techniques of information and communication technology tools for various applications (Ting 2015).

Besides straightforward courses, other movements and initiatives are also contributing to spread data literacy in higher education. One of such proponents of data literacy is the Open Data movement. Data and content are deemed "open" when they "can be freely used, modified, and shared by anyone for any purpose" (Open Knowledge Foundation 2020). The idea of open data dated back to the 1990s (Chignard 2013). Nowadays, open data is a wide-ranging topic applicable to culture, science, finance, statistics, weather, environment, and many more. On a larger scale, in recent years, modern administrations have been embracing the Open Government Data philosophy (Ubaldi 2013). Open data has three key features: 1) availability and access, 2) reuse and redistribution, and 3) universal participation. These features explain why the Open Data movement is relevant to Data Literacy. It requires data competences from both sides. Institutions, organizations, and companies must know how to share their reusable and interoperable data. End-users have to know how to access, analyze, and interpret the data. In fact, the Open Knowledge Foundation is a vocal advocate of data literacy (Open Knowledge Foundation 2020). The Open Data movement is especially beneficial for education in general and higher education in particular. The more datasets there are, the more practice opportunities there are for students and researchers. However, alongside the positive aspects, there are also concerns regarding Open Data and data openness. This topic will be discussed further in Section 3.1.

In line with the Open Data principles, the Open Educational Resources (OER) initiatives are also a strong proponent of data literacy. "OER are teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or re-purposing by others" (Atkins, Brown & Hammond 2007). OER's creators often use respective Creative Commons (CC) licenses to indicate what others could and could not do with the resources. In particular, CC0, CC BY, and CC BY-SA are Creative Commons licenses that could be considered truly open for all public usages. "OER can include full courses/programmes, course materials, modules,

student guides, teaching notes, textbooks, research articles, videos, assessment tools and instruments, interactive materials such as simulations and role plays, databases, software, apps (including mobile apps) and any other educationally useful materials" (UNESCO 2015). The Hamburg Open Online University (HOOC) in Germany is a pioneer in this area. Eight social institutions actively participate in this inter-university network, including the Ministry of Science, Research and Equality, five state universities in Hamburg, the University Medical Center Hamburg-Eppendorf (UKE) and the Multimedia Kontor Hamburg (HOOC 2020).

The Open Data movement and OER are frontline initiatives advocating for data literacy. Despite such movements still being relatively modest in quantity and resonance, their moderate success gives hope for the future of data literacy. Training programs in higher education have to keep up with the rapid technological evolution to bring substantial added value. Ridsdale et al. (2015) concluded that "best practices for teaching data literacy education include collaboration between educators, organizations, and institutions to ensure goals are being met by all stakeholders; diverse and creative teaching approaches and environment including the effective use of technology; successive/iterative learning with complementary skills integrated (e.g. project--based learning); emphasizing mechanics in addition to concepts (i.e. practical, hands on learning); and increasing engagement with the content by using real world data."

All in all, as data is gaining importance in public and private life, data literacy is gaining momentum in the educational landscape. The need for data literacy is growing due to the growing complexity of new technological challenges. Though the selection of training programs has yet to be considered comprehensive, observations show that institutions, universities, and initiatives are making efforts to bring data literacy to higher education.

3 Data ethics

3.1 Ethics and data

3.1.1 Philosophical basis

Ethics, or moral philosophy, is not a new subject. As early as around the fourth century BC in the Western world, there had been three major philosophical pillars making up the ancient Greek philosophy: Socrates, Plato, and Aristotle. Derived from the Greek "ethos", meaning "custom" or "habit", ethics is the branch of philosophy that deals with determining the proper course of action for humans. Ethics is about "doing the right thing; the philosophy behind it is about determining what those right things are, in a way that benefits the individual and society at large in a fair, just, and kind manner" (Boone 2017). In the East, Confucianism and Buddhism dominated the philosophical discourse. Despite its simple definition, ethics is a controversial topic, and different schools of philosophy view ethics differently. For example, the trolley dilemma (Thomson 1985), the prisoner's dilemma (Cunningham 1967), or the liar paradox (Rabaté 2008) sparked how different schools of philosophy justify a decision differently.

The schools of philosophy concerning ethics could be attributed to three main categories depending on the philosophical question. In particular, meta-ethics ask "What does 'right' mean?", normative ethics ask "How should a person act?", and applied ethics ask "How do we act according to our moral principles?". Despite differences in interpretations, each of the ethical theories could contribute to the definition of ethics as a multi-faceted concept.

Meta-ethics investigate a broad range of questions and puzzles to inquire about the metaphysical, epistemological, semantic, psychological, presuppositions, and commitments of moral thought, talk, and practice (Sayre-McCord 2012). With the answers to those questions, meta-ethics explore the connection between values, reasons for actions, human motivation, and ways to support or defend the nature of ethical properties, statements, attitudes, and judgments. For instance, questions that fall within the meta-ethics' domain could be: "Is morality more a matter of taste than truth? Are moral standards culturally relative? Are there moral facts? If there are moral facts, what is their origin? How is it that they set an appropriate standard for our behavior? How might moral facts be related to other facts (about psychology, happiness, human conventions…)?" (Sayre-McCord 2012).

Among the branches of normative ethics, it is notable to mention virtue ethics, consequentialism, and deontology. First, virtue ethics originated from the ancient Greek philosopher Aristotle. Virtue theory dictates that the good habits that humans form will lead to happiness and a good life. Therefore, the virtuous person is objected to practical wisdom, which means that s/he should strive towards making clear choices that will help shape the person s/he wants to become. Second, according to consequentialism and its derivative utilitarianism, outcomes are the ultimate merit. An action should lead to the best possible outcome, no matter the intention behind it. There is no set measurement scale for good outcomes, as well as it is subjective to determine the beneficent of these good outcomes. Nevertheless, consequentialists make clear that everyone matters equally. Third, deontology–the rival school of thought to consequentialism–stresses the importance of the intention. Deontologists say that it is imperative to do the right thing, no matter the consequences. The German philosopher Immanuel Kant was a deontologist. He believed that the ability to reason gives humans moral status and that this moral worth is intrinsic.

Besides normative ethics, during the Age of Enlightenment, a book by the Genevan philosopher Jean-Jacques Rousseau inspired the genesis of the social contract theory which then became a branch of ethics and was regarded as the leading doctrine of political legitimacy of the epoch (1762/1950). It concerns the power dynamic between the state and its citizens. Contractualism seeks to justify both parties' rights and responsibilities and claims that a social contract is the basis of legitimate authority. A legitimate state is one that its people granted consent to exercise power.

Despite their differences, no philosophical school of thought is entirely right or wrong. Historically, schools of philosophy are not separated from each other, but they instead make up a patchwork system where successors either complement or contest the opinions of their predecessors. As a whole, they contribute valuable insights to the grand scheme of how humans should live. Regardless of one's philosophical belief, studying and applying ethics remains essential. It teaches a person to be a good citizen, thus allowing them to exist in harmony with other beings. "Ethics in the broadest sense refers to the concern that humans have always had for figuring out how best to live" (Vallor 2019). Applied ethics can be found in all aspects of life: political, personal, and professional. The search term "ethics" retrieved 1129 documents from the Stanford Encyclopedia of Philosophy

(2020). Humans accumulate knowledge; thus, the discussion on how to apply this knowledge in the best way possible is a continuous process.

3.1.2 The three conceptual axes of data ethics

The teaching of ethics represents the branch of applied ethics which has been taught in more traditional disciplines for a long time, e.g., medicine, law, and engineering. In comparison to these disciplines, technology ethics has just emerged in recent decades. Since 2014, one of the world's largest technical professional organizations, the Institute for Electrical and Electronics Engineers (IEEE), has been including technology into its annual international conferences on ethics in science and engineering. Today, IEEE has an entire division devoted just to technology ethics called the IEEE TechEthics program, i.e., the ethical and societal impacts of the technologies. As with most things in life, technology is never ethically neutral. Every technology reflects the values and decisions of its creators. What might be a benefit for a party could bring a disadvantage to others. With the rampant technological development, the world cannot lose sight of ethical issues; otherwise, humanity will more frequently face incidents like the Snowden revelations, or the Cambridge Analytica scandal.

Data ethics find itself at the crossroad of technology ethics and data literacy. Data ethics concern the alignment between moral principles, the understanding of data, and data-related decision making. As illustrated in Section 2.1, data concerns everybody. From the collection to the evaluation of data, ethical issues continuously emerge. There are three main aspects of data ethics: the ethics of data, the ethics of algorithms, and the ethics of practices. The ethics of data concern how data is generated, recorded, and shared. The ethics of algorithms concern how AI, machine learning, and robots interpret data. The ethics of practice concerns devising responsible innovation and professional codes to guide this emerging science (Floridi & Taddeo 2016).

During the discussion on the ethics of data, a recurring keyword is "big data". Big data refers to the massive amounts of data collected, stored, and analyzed by anyone who has access to data sources. Most commonly, they are governments, companies, and services. Historically, people have been collecting data in analog forms, long before the emergence of digital technology. Since the invention of computers and digital storage, data has always been collected and analyzed. However, it is big data that holds a critical factor in the way technology has been

transforming the world. Thanks to big data technology, vast amounts of collected data can be analyzed to generate value (Watson 2014). Data is considered "a new source of immense economic and social value" (Emrouznejad & Charles 2019).

Similar to "data literacy," the definition for the term "big data" is ever-evolving. On an elemental level, big data can be defined using the 3Vs approach: volume, velocity, and variety (Zikopoulos et al. 2012). Big data refers to data that is huge in volume, high in velocity, and diverse in type. The term "big data" can be traced back to the mid-1990 when John Mashey, retired former Chief Scientist at Silicon Graphics, described the handling and analysis of massive datasets (Diebold 2012). Initially, very few people used the term "big data," neither in the academic world nor in industries. As the 2010s decade came around the corner, amidst worldwide technological upheavals, "big data" became a buzzword frequently used in business and the popular media (Kitchin 2014). Discussion on big data started picking up as renowned publications such as The New York Times (Lohr 2012), Financial Times (Taylor 2012), Science (Mervis 2012), and Nature (Marx 2013) frequently reported on the subject.

Big data can bring both benefits and harms to society. Better human understanding, social, institutional, economic efficiency, predictive accuracy, and personalization technologies reflect these benefits. Despite these achievements, unethical big data practice has and will potentially cause more harm to privacy and security, fairness and justice, transparency, and the autonomy of the individual. "Even when a data practice is legal, it may not be ethical, and unethical data practices can result in significant harm and reputational damage to users, companies, and data practitioners alike" (Vallor 2019).

Two keywords that often come up with the collection and sharing of data are trust and transparency (Floridi & Taddeo 2016). While the lack of trust and transparency challenges the data collection, overstating trust and transparency will challenge data sharing. The ethics of data consider both sides of the spectrum and strives towards finding a balanced interplay of skepticism and openness in data practices. While initiatives like the Open Data movement promotes transparency, researchers have repeatedly misinterpreted the openness of data. In this regard, research that obtains data from social media platforms are proven to be problematic (Zimmer 2010, 2018; Hunter & Evans 2016). On the one hand, it is the researcher's responsibility to acknowledge that data openness does not necessarily imply consent to data collection and exploitation for other purposes.

On the other hand, the recurring debate on research using social media data also reveals gaps in governance and regulations.

Another axis of research entails the ethics of algorithms, which discuss the way AI, machine learning, and robots interpret data. Advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, and natural language processing co-exist with the increase of data generated. Algorithms were designed as a delegate for making decisions, but they have yet to reach their maturity. Unintended behaviors, lack of foresight, the difficulty of oversight, distributed responsibility, various kinds of risks, are among other problems that have been confronting our society. The gap between algorithms' design and operation (Mittelstadt et al. 2016) is the center of discussion around the ethics of algorithms. In 2016, the car company Tesla brought onto the market a new generation of the Model S and claimed to have equipped it with "full self-driving hardware" (Tesla 2016a). The product launch took place only four months after a fatal accident caused by a flaw in the car's autopilot mode, where the algorithm could not distinguish a white truck obstructing the path (Tesla 2016b). In Tesla's defense, the fatality rate correlating with Tesla models is considerably lower than the average fatality rate among all vehicles in the US or worldwide.

Nevertheless, while Tesla could justify its algorithm by blaming on immaturity, it shows that ethics requires more than "good intentions". Without a conscious mindset for maximizing benefits and minimizing harms, many bad choices could still be made by persons who meant no harm (Vallor 2019). The shift in Google's manifesto from "Don't be evil" to "Do the right thing" aptly supports this point. It is possible to avoid making an unethical decision, yet this avoidance could still come at the cost of others. To "do the right thing", as Google has put it, is "to follow the law, act honorably, and treat each other with respect" (Alphabet Inc. n.d.). Whereas the previous mantra prohibits conflict of interests, the latter directly requires that the decision-makers consider all stakeholders to ensure that the outcome will benefit everyone. There is a transition from compliance to conscience.

The remaining axis of data ethics concerns the ethics of practices, focusing on devising responsible innovation and professional codes to guide the emerging field of data science. The development of new data products involves many parties, e.g., the programmer, the sponsor, the UX/UI designer, the technical support. Who should assume responsibility in worse cases, and if so, to what

extent? It is dangerous to separate science and technology developments from ethical evaluations (Leonelli 2016) unless technology is not designed to serve humans. Not only is a reflection on the potential impacts a critical task of a researcher, but it should also become an integral part of scientific work. Furthermore, if an ethical evaluation is lacking from the beginning of development, the chances are that a reflection afterward could only point out the problems, but it would be too late to start over.

In short, ethics is a philosophical topic in which humans have been engaging since ancient times. Within the vast field of philosophy, ethics finds itself in the discussion of human morality and concerns in particular how humans should live. Whereas ethics has been a curriculum part of more traditional majors such as medicine, law, and engineering for hundreds of years, technology ethics has just been on the horizon since recent decades. On the foundation of established philosophical schools of thought, ethics has been evolving to accommodate the changes in modern society, especially in the hyperconnected world enabled by modern technology. At the crossroad of technology and ethics, data ethics was born. Three conceptual axes constitute data ethics. They overlook the ethics of data, the ethics of algorithms, and the ethics of practices. Each aspect associates with different ways that the use of data is affecting humankind.

3.2 Prerequisites for ethical data practice

3.2.1 Ethical oversight

Among current literature on data ethics, the keyword that often comes up is an oversight. Oversight of data ethics must include all three aspects: data, algorithm, and practices. Despite its frequent association with authorities, oversight takes place not only via guiding frameworks and legal regulations but also via individual practice. Therefore, the discussion about ethical oversight must include both policymakers and end-users and data professionals to facilitate data democracy.

On the personal level, ethical oversight ensures individual data control, transparency, equality, autonomy, and accountability (Transberg et al. 2018). As technology is created to serve humans, the human should be at the center of technology design. The data that technologies rely on are generated by humans. Therefore, it is comprehensible that individuals should be able to exercise their

right over their own data. However, individual data control requires an important prerequisite–data literacy. A person who is data literate is able to recognize lack of transparency and fairness in data collection. S/he is able to demand autonomy when it is not provided. Moreover, s/he is able to determine if a data source is reliable, and if the data collecting method is sustainable, hence accountability.

On the professional level, each discipline needs its own code of conduct. Regardless of the differences in professions, one must make sure that his actions are not only compliant, but also cause no harms to others. Especially in the rapid expanding field of AI and machine learning, where there is much less control over the outcomes, even more thoughts must be put into the drafting of projects. In the recent years, at least 84 public-private initiatives have produced statements describing high-level principles, values, and other tenets to guide the ethical development, deployment, and governance of AI (Mittelstadt 2019). However, Mittelstadt also criticised that this is not automatically a good sign, as "AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms" (2019). Governance of AI in general and algorithms in particular must monitor the incentives of stakeholders in AI development.

On the political level, oversight poses questions about power, justice, and responsibility. Policymakers bear the responsibility to design legal frameworks that protect both individuals and government alike. Then again, revelations on mass surveillance programs from governments all over the world cannot help but raise questions about the motive behind these political arrangements. There is great demand for active citizen participation in the political discourse of data ethics.

Ethical oversight is not a black and white topic. There is a fine line between ethical practice and skepticism. User's trust is easy to lose because it concerns private data that can cause great harm if exploited. Some echo chambers have been forming in public discourse, especially on social networking platforms (), where the algorithms tend to suggest content based on previous interactions of the users. In the case of public opinion on artificial intelligence, Tegmark described three main camps with entirely different views as Digital Utopians, Techno-Skeptics, and The Beneficial-AI; among the three, he identified himself with the latter group (2018). According to Tegmark, artificial intelligence can be extremely

beneficial for humankind, if and only if humans can ensure AI robustness that meets the standards of verification, validation, control, and security. Tegmark is one voice among many other influential proponents of the use of data and big data technologies. The Future of Life Institute (FLI), where he is a founding member, is a group of data experts who believe in a future where humankind will prosper with AI. FLI founding members include well-known figures such as the co-founder of Skype Jaan Tallinn, DeepMind research scientist Viktoriya Krakovna, and SpaceX and Tesla Motors founder Elon Musk.

Meanwhile, Techno-Skeptics are also vocal about their concerns, and the echo is getting louder (Achenbach 2015). Lanier (2011), Keen (2015), Barrat (2015), and Taylor (2015) are among others resonating voices who strongly criticize the erosion of privacy, technological dependency, as well as gender and racial disparities as a consequence of uneven adoption among different groups of people. Thus, it shows that even the body of experts has not been able to obtain a unifying view on the ethics of data, let alone the general public. There is a need for more direct exchange between these camps to foster constructive discussion on ethical oversight. exchange between these camps to foster constructive discussion on ethical oversight.

3.2.2 Current practical implementations

To possess data ethics is to acknowledge one's moral values, set an intention before making any effort, and frequently review its potential consequences. Each person will know what s/he does not want to happen with the data that s/he generates. To be ethical in practice is to be aware that if something should not happen to oneself, it should not happen to others (Transberg et al. 2018).

In a 2019 report, the British Royal Society identified the need to develop data science as a profession. Thereby, data scientists can adhere to a professional framework with "shared codes of practice, including appropriate governance of data collection and use and ethics training is an important short-term goal. In the longer term, professional bodies such as the British Computer Society and the Royal Statistical Society, could work with employers and universities and identify the skills needed for data scientists and consider how to address accreditation to ensure that students and professionals can be confident in the quality of new courses" (Royal Society 2019b).

Globally, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is a platform where individuals or representatives of organizations can exchange their standpoints, thus collectively prioritize ethical considerations in the design of autonomous systems. For two years from 2017 to 2019, over seven hundred volunteers helped to create Ethically Aligned Design (EAD), a comprehensive report that "combines a conceptual framework addressing universal human values, data agency, and technical dependability with a set of principles to guide autonomous and intelligent systems creators and users through a comprehensive set of recommendations" (IEEE 2019).

With some new modern laws of data protection such as the European General Data Protection Regulation (GDPR) or the UK Data Protection Act as an adaptation of the GDPR, it became possible for an end-user to discover who collects and stores their data. The GDPR entered into force on 24 May 2016 and has been effective since 25 May 2018. It aims to protect natural persons regarding the processing of personal data and the free movement of such data. With its Europe-wide legality, the GDPR standardizes fragmented national standards.

Even though the GDPR also impacted non-European businesses, there has not been an equivalent in the US and other countries. The GDPR and the UK Data Protection Act apply to personal data stored within European and British territories. For personal data generated inside Europe to be transferred to, processed, and stored in the US, where data protection standards are significantly lower, companies have to comply with the US-EU Privacy Shield Framework. Designed by the US Department of Commerce and the European Commission in 2016, Privacy Shield is a critical enabler for transatlantic commerce. Its predecessor from 2000 to 2015 was the US-EU Safe Harbor Framework, which could not fulfill the standards of the 2016 European GDPR. On 16 July 2020, following the Schrems II case, the Court of Justice of the European Union ruled to invalidate the US-EU Privacy Shield[1]. Companies with activities within the European Union and the European Economic Area must now comply with the Standard Contractual Clause (SCC) to transfer their data to servers outside of these areas. Even though the transition from Privacy Shield-based to SCC-based

[1] Judgment of the Court (Grand Chamber), 16 July 2020, Case C-311/18, Judgment ECLI:EU:C:2020:559

data handling could cause companies time, labor, and money, the decision to safeguard personal data generated inside Europe is definitely an ethical move.

For professionals, there is a huge selection of toolkits for professionals and companies who want to make their practice more ethical. In the USA, the nonprofit Center for Humane Technology (CHT) was launched in 2018 by former Google design ethicist Tristan Harris. The main works of CHT consist of educating the public, informing policy change, and supporting technologists. CHT identified that the problem with tech platforms is their attention extraction economic model, where these platforms profit from human emotional contagion (Kramer, Guillory & Hancock 2014). Hence, CHT's mission is to reimagine technology infrastructure and business models that actually align with humanity's best interest. CHT created a worksheet called Humane Design Guide for product designers. By completing this worksheet, creators will be made aware of ethical decision-making, including potential applications for data collection, data sharing, and algorithms and the consequences.

Developed by the Institute for the Future and Omidyar Network, EthicalOS is a toolkit for technology companies on how to be more ethical in operation. In Silicon Valley, Integrate.ai offers a solution based on artificial intelligence to manage customer relations using signals instead of personal data. In the same field, early 2020, the Singaporean Personal Data Protection Commission (PDPC) issued the second edition of the Model Artificial Intelligence Governance Framework. It is a set of guidelines on AI implementation for organizations. On the corporate side, big technology players such as Google, Apple, Microsoft, SAP, Accenture, and Twitter also published principles pledging ethical practice. In practice, there is room for discussions if these companies' practices have lived up to their own standards.

For end-users, the number of privacy-first solutions and services is steadily on the rise. Most notably, there are Virtual Private Networks (VPN) providers that allow users to mask their actual geographical locations. Privacy-concerned Internet users may know the Tor Browser, a web browser developed by The Tor Project whose mission is to protect people's identity online. Sharing the same dedication as The Tor Project, end-to-end instant messaging services such as Telegram, Signal, and Threema allow people to chat without the fear of being spied upon freely. For the more traditional way of communication, there is the Swiss-based end-to-end encrypted email service ProtonMail or the open-source email client

Thunderbird. Competing with Google, the search engine DuckDuckGo is pledging to help users “take back their privacy.” Similarly, the decentralized social network Diaspora envisions an online social world where users can control their personal data.

The outstanding issue with digital products and services that users commonly encounter lies in lengthy Terms and Conditions (T&C), typically with small prints and legal jargon. Even though T&C are legal and come with virtually every digital service, they are not drafted with inexpert users in mind. Moreover, users tend to skip the T&C in order to use the services right away. This kind of practice challenges transparency; it has the potential to jeopardize users’ awareness about their rights to privacy protection.

All in all, whereas data ethics overlooks the potential advantages and disadvantages data and the use of data can bring, ethical oversight is gaining in importance to promote good practices. Worldwide, there have been positive efforts in commercial products as well as in legislation. Nevertheless, not all countries and regions are on the same level regarding good practices. This disparity could lead to misalignment and even contradicting mindsets in practices. The world needs more good practices, especially on a global scale. More ethical, institutions, professional societies, universities, and companies should strive towards better collaboration. Regardless of emerging challenges, good practices need to keep up with technological advances.

4 Analysis of data ethics in the context of data literacy training for higher education

4.1 Research focus on data ethics

Current happenings show that data literacy is gaining importance as an academic competence. Along with the efforts in establishing data literacy as a basic academic competence, it is crucial to consider its ethical aspect. While data literacy promotes the necessary skill to understand and use data in a critical manner, data ethics is a step further that requires this manner to be ethical. A public survey from Hochschulforum Digitalisierung (HFD), a joint initiative of Centrum für Hochschulentwicklung (CHE), Hochschulrektorenkonferenz (HRK) and the Stifterverband, found out that competence on data ethics plays a significant role in society (Schüller, Busch & Hindinger 2019). To illustrate, data projects that concern sensitive personal data often require the professional support of data protection experts and data ethicists, who are qualified to perform data security measures such as pseudonymization. Similarly, other scientific research on data literacy has repeatedly mentioned data ethics as a crucial aspect of data literacy, mostly as the competence to work with data and apply data (Manduca & Mogk 2002; Maybee & Zilinski 2015; Ridsdale et al. 2015).

Worldwide, there are various researchers and research groups that are actively contributing to exploring the topic of teaching data literacy in higher education in a variety of disciplines, for instance, social sciences, life sciences, and teacher education (Prado & Marzal 2013; Carlson & Bracke 2015; Dunlap & Piro 2016; Gibson & Mourad 2016; Wolff et al. 2016; Ming & Hui 2018). Regarding training on data ethics, even though some researchers have been vocal about the importance of this topic (Floridi & Taddeo 2016; Mittelstadt et al. 2016), there are fewer empirical findings (Manduca & Mogk 2002; Maybee et al. 2015; Ridsdale et al. 2015; Schüller, Busch and Hindinger 2019).

The competency to use data ethically is listed in ILSHE, “Information Literacy Standards for Science and Engineering/Technology” (ILSSET) and “Information Literacy Competency Standards for Journalism Students and Professionals” (ILCSJ) by the ACRL. However, training on data ethics remains scarce and mostly optional, if at all offered for higher education purposes, even though it would benefit anyone regardless of socio-economic as well as educational background (Royal Society 2019b). It is probably because the topic has not yet

gained importance in the public consciousness and that people with a higher level of education (e.g., data scientists) are the first to be confronted with the topic. Chapter 3 illuminated that data ethics is an ambiguous subject. On the transition towards a data-literate and data-ethical world, there is a need for more research on the implications of data ethics and potentially practical education approaches.

From 2013 until mid-2019, there has been a sharp rise in UK job-listings for 'Data Scientists and Advanced Analysts' (+231%) driven predominately by increased numbers of vacancies for Data Scientists (+1287%) and Data Engineers (+452%) (Royal Society 2019b). "Data Scientists are highly trained and curious professionals with a taste for solving hard problems and a high level of education (often Ph.D.) in analytical areas such as statistics, operational research, computer science and mathematics" (Watson 2014). Data Scientists often come from a mathematical or statistical background. However, according to Emma McCoy, Professor of Statistics and Vice Dean of Education at the Faculty of Natural Sciences, Imperial College London, the nature of skills is changing as well, mainly due to the increase of Machine Learning prediction tools based on historical data, an understanding of fairness and ethics would be required from Data Scientists (Royal Society 2019b) alongside discipline-specific skills.

When it comes to handling data, there are different expectations according to each user. In the academic field, research activities often require ethical data practices. However, the scope is much more significant. Many majors, especially but not exclusively programming-oriented ones, prepare students for the professional world, where the decision-making process should base on ethical grounds (Maybee et al. 2015). Higher education provides the stepping stone to the professional world, including but not limited to professions in data science. From the beginning, assignments and projects sensitize students to handle data, e.g., survey results or gathering information for essays or presentations. Students are young adults who are forming their worldview and navigating their moral compass. It is at this crucial stage that they should be guided towards ethical values. Therefore, data ethics must be taught more widely across the globe.

This work will concentrate on the teaching of data ethics. However, instead of solely focusing on ethics in data science, the main question that concerns this research is: Regardless of the discipline, how is data ethics being taught in higher education? This research aims to bridge this gap by generating new findings on data ethics that would contribute to the education efforts towards widespread data

literacy as a whole scheme. As many training materials on data ethics focus on higher education, it is suitable first to expand research in this area and, eventually, to other fields in terms of a more general audience. Generating new findings on data ethics would contribute to the education efforts towards broad data literacy as a whole scheme. The main interest of this research is training materials on data ethics. Therefore, it will prioritize these materials. As data ethics is considered a small aspect of data literacy education, the research will also consider materials on data literacy that include a portion of data ethics.

4.2 Considerations for a qualitative method

Quantitative research follows a deductive research process, starting with hypotheses that can be tested, asking questions about frequency and quantity. In contrast, qualitative research follows an inductive research process to gather explanations and meaning, asking questions about motive and mechanism (Coleman & O'Connor, 2007). Qualitative methods explore substantive areas about which little is known or about which much is known to gain new understandings (Stern, 1980). Additionally, qualitative methods can obtain intricate details about phenomena such as feelings, thought processes, and emotions that are difficult to extract or learn about (Corbin & Strauss 1998). A qualitative analysis of existing training syllabi on data ethics in higher education would help pinpoint common training objectives, training methods, and potential criteria for evaluating ethical data practices.

Ethics is an emotional topic, one that concerns the ontological existence of human beings. Because teaching methods may vary, this work does not aim to describe the teaching of data ethics, rather conceptualize the frameworks of teaching data ethics. Whereas a quantitative survey can provide insight into the learning process, as training on data ethics has been scarce and mostly disciplinary, it is a challenge to collect enough data to represent the results. Then again, qualitative research methods are great tools for gaining new insights and substantiate ambiguous topics. For the particular purpose of this research, the grounded theory methodology was chosen in the hope that it would provide the right tools first for laying the groundwork for future findings on the teaching of data ethics with the extraction of pedagogical concepts.

First introduced in 1967 by Glaser and Strauss, grounded theory methodology is an inductive approach to research which emphasizes the process of finding new theories and phenomena over testing existing hypotheses. It was initially designed

for sociology, as both Barney Glaser and Anselm Strauss were active researchers and educators in this field. Later on, as Strauss continued to apply and teach the grounded theory approach, he and Juliet M. Corbin collaborated to develop the Straussian branch of the original approach. Glaser remained faithful to the more positivist functionalist sociology of Merton and Lazarsfeld at Columbia, with whom he originally trained. Meanwhile, Strauss became increasingly constructionist in his thinking and remained an interpretive symbolic interactionist all his life, at least as devoted to developing interactionist theory as he was to methods (Clarke 2019).

The theory harvested with the grounded theory approach is "derived from data, systematically gathered and analyzed through the research process" (Corbin & Strauss 1998). A theory is "a set of well-developed concepts related through statements of relationship, which together constitute an integrated framework that can be used to explain or predict phenomena" (Corbin & Strauss 1998). Although it is not necessarily the only end goal of doing research, a theory has a vital role in science (Strauss, 1995 via Corbin & Strauss 1998).

The design of grounded theory procedures facilitates the development of concepts that explain the theories of social phenomena under study (Corbin and Strauss 1990). The inductive nature of the grounded theory approach assists in a research attitude that considers the point of view of the subject under study, in this case, of the educators who designed the curricula. By choosing this approach, the research can concentrate on a thorough analysis of training curricula to extract well-integrated concepts that would best explain how data ethics is being taught in higher education, rather than imposing set criteria on a somewhat less discussed topic. The researcher begins her research without any preconceived theories.

Other researchers have been using the grounded theory approach to research data literacy. For instance, Maybee et al. (2015) conducted an inductive analysis of syllabi on information literacy and data information literacy at the nutrition science and political science faculties at Purdue University. The findings revealed the relationships between data literacy and other concepts in the syllabi and provided a comparison of training goals at these two faculties.

4.3 Methodological approach

4.3.1 Data base

After reviewing the literature on data literacy and data ethics and becoming more sensitized to the topics of the research in Chapter 2 and 3, the researcher started collecting data. Due to the nature of the research objectives, the analysis will mostly deal with secondary data, i.e., curricula, frameworks, published books, journals. Materials were accessed through the Google search engine, DuckDuckGo search engine, Ecosia search engine, Google Scholar, Elsevier, ResearchGate, Academia, and other indexes from May through July 2020. Resources were identified by using the keywords "data literacy training," "data literacy curriculum," "teaching data literacy," "data literacy course," "data ethics," "teaching data ethics," "data ethics course," "data ethics curriculum," "data ethics framework" as keywords. The researcher then narrowed the search results by picking out sources which concern the relevant target–undergraduate and graduate students. The query resulted in 27 sources of varying lengths and formats (see Appendices), including university course module manual, competence frameworks from institutions, and guidelines from professional societies. While there are syllabuses that dedicated data ethics explicitly, there are also general competence frameworks that only mention this topic briefly without going into the details.

Therefore, to achieve comparability between sources, the researcher had refined the selection of sources for the qualitative analysis to only materials that discussed data ethics training in greater detail. Furthermore, due to differences in detail degree and format, some documents are taken fully into consideration, while only parts of other documents are chosen for the analysis. Typically, the chosen parts are the introductory parts, which include essential information such as learning goals, course content, and assessment methods.

4.2.2 Methodical procedures

The following procedures followed the Straussian branch of grounded theory methodology, which Anselm Strauss and Juliet Corbin further developed on the grounds of the original 1967 methodology from Glaser and Strauss. The undertaken techniques and procedures for developing grounded theory followed Strauss and Corbin's recommendations in their 1998 publication on the basics of qualitative research. This publication served as a reference point so that the

following procedures could guide the analysis while granting the analyst the flexibility to adjust the sequence of steps along the way.

As with researches applying the grounded theory methodology, the research started with an open coding phase. Open coding consists of breaking down the curricula data into discrete parts, close examining, and comparing for similarities and differences. Analysis of a word, phrase or sentence comprises of scanning a document, or at least a few pages, and afterwards returning to focus on a word or phrase that the analyst considers to be significant and analytically interesting. Then the analyst starts listing all the possible meanings of the word which comes to mind. The next step is to group events, happenings, objects, and actions/interactions that were found to be conceptually similar or related in meaning under more abstract concepts termed "categories." Even though asking questions and performing comparative analysis are recurring analytic tasks and used systematically throughout the analysis, these techniques were especially useful during the beginning phase. Via open coding and the discovery of categories, concepts started to develop.

In case a category was not fully developed, there were various ways to densify. First, the researcher could resort to theoretical sampling. Theoretical sampling is the act of asking questions to compare concepts derived from the evolving theory with other concepts to discover the variations among these concepts and further expand the properties and dimensions of the underdeveloped categories. Second, the analyst could expand a category by using the paradigm to determine whether a category denotes a condition, an action/interaction, or a consequence. The paradigm can make the nature of a category more apparent. Third, for an underdeveloped category that denotes an action/interaction, it is worthwhile to observe whether a transformation of the process took place and how the actions/interactions evolve according to changes in their contextual factors. Even if an action/interaction tends to be a routine, it is still beneficial to clarify which factors provide the conditions for it to stay repeated. The theory could either spin around the process or do not correlate with the process. Nevertheless, the process provides insight into the evolution of actions/interactions over time and space (Corbin & Strauss 1998).

Parallel to creating categories, the analyst also created subcategories and determine the relationship between a category and its subcategories through axial coding. Axial coding is "the process of relating categories to their subcategories,

termed ‘axial’ because coding occurs around the axis of a category, linking categories at the level of properties and dimensions” (Corbin & Strauss 1998). Occasionally, the researcher performed a microanalysis (microscopic analysis) of coded data. Microanalysis consists of “detailed line-by-line analysis at the beginning of a study to generate initial categories (with their properties and dimensions) and to suggest relationships among categories” (Corbin & Strauss 1998). By performing microanalysis, an analyst could recognize “vivo concepts”– phrases used repeatedly and so representing events that are probably important. Once categories are established, analysis becomes more focused on filling out those categories and verifying relationships. Microanalysis methodically combines open coding and axial coding. In the end, the most priority is to figure out the relationships between categories.

Open coding, axial coding, microanalysis, theoretical sampling, paradigmatic observation, and process analysis are continuous procedures. Research design can only be best applied to statistical sampling, whereas with theory building, the researcher must continuously go back and forth between the data to gather new findings. Towards the end of the analysis, the analyst started to integrate and refine the theory by choosing a central category. This central category, which Strauss also referred to as the “core category” (1987), should meet a set of criteria. For example, the central category must be related to as many other categories and their properties as possible; the indicators pointing to the central category frequently appear in the data; the central category can be easily related to other data. The central category opens the doors to building the maximum variation in terms such as dimensions, properties, conditions, consequences, and strategies (Corbin & Strauss 1998).

Along every step of the analysis, the analyst created memos to document their progress, thought process, and observations. These memos are equivalent to the field notes of social scientists. They reflect on how the actual analysis took place. Using these memos as reminders, the analyst could write up their research results in the most detailed manner possible. Another way to demonstrate the research progress is to utilize diagrams. Visualizations can emphasize structure, process, and relationships exceptionally well.

During qualitative research, the analyst handled different formats of data and generated many memos. To optimize the workflow, the program MaxQDA was chosen as the analytic tool. This way, the analyst could have a better overview of

research materials and the analytic process. Besides, the use of MaxQDA helped increase the transferability and portability of research data, thus contributing to the reproducibility of analysis results.

5 Results

5.1 The beginning phase of the analysis

The main research objects of this inductive analysis are data literacy curricula/frameworks that address training on data ethics for undergraduate and graduate students. As shown in Table 1, there were nine from 27 collected sources used for the inductive analysis. These materials were chosen on the grounds of their direct reference to data ethics in the context of data literacy training.

Table 1: List of sources chosen for the analysis

Nr.	Name	Provider/Country	Year	Assessment methods	License for re-use
1	Data Science Ethics	University of Michigan (USA)	2020	Video lectures Graded quiz Written discussion Peer-graded assignment	All rights reserved
2	An Introduction to Data Ethics	Markkula Center for Applied Ethics at Santa Clara University (USA)	2018	Questions Case studies	Free with permission to use
3	Curriculum framework for undergraduate and graduate students in science, health sciences, and engineering programs (Module 4, 5, 6)	Lamar Soutter Library, University of Massachusetts Medical School and the George C. Gordon Library, Worcester Polytechnic Institute (USA)	2012	Questions Case studies Written discussion	CC BY-NC-SA 3.0

4	Instructor Guide to Flipped Data Management Course	University of Minnesota (USA)	2014	Develop and implement a data management plan	CC BY-NC 4.0
5	Data Management Expert Guide (Introduction & Chapter 5)	CESSDA Training Working Group (Consortium of European Social Science Data Archives) (EU)	2017–2019	Questions for ethical review (self-assessment or formal)	CC BY-SA 4.0
6	Guide to Developing a Data Protection Management Program	Personal Data Protection Commission (Singapore)	2019	Not applicable	All rights reserved
7.1	New England Collaborative Data Management Curriculum (Module 1)	UMass Medical School, Lamar Soutter Library (USA)	Since 2012	Case studies Questions	CC BY-NC-SA 4.0
7.2	New England Collaborative Data Management Curriculum (Module 5)	UMass Medical School, Lamar Soutter Library (USA)	Since 2012	Case studies Questions	CC BY-NC-SA 4.0
8	Learn to Analyze Educational Data and Improve your Blended and Online Teaching Massive Open Online Course	Learn2Analyze — An Academia-Industry Knowledge Alliance for enhancing Online Training	2020	Graded quiz	All rights reserved

	(MOOC)	Professionals' (Instructional Designers and e-Trainers) (EU)			
9	Opinion of the Data Ethics Commission (Executive Summary)	Data Ethics Commission (Germany)	2019	Not applicable	All rights reserved

A brief overview of the chosen sources reveals an assortment of training materials, ranging from university module syllabuses, course instructor guides, to expert guides from professional societies. Moreover, the sources also vary in origin countries, publishing year, assessment methods, and re-use licenses. Due to these differences, it is inevitable that the scope of these materials also differs significantly from each other. For example, while source #2 generally addresses the ethics of students in the role as data practitioners, source #3 specifically addresses competences in data management among students in science, health sciences, and engineering programs, where module 4, 5, 6 gave details of legal and ethical considerations for research data. Based on this observation, the analysis required prioritization of which parts of a source to consider. Therefore, with sources whose focus is specifically data ethics, the whole document was taken into account; with sources that were more general or included contents irrelevant to data ethics, only the parts dedicated to data ethics were taken into account. The parts chosen for analysis were also listed in Table 1.

After the open coding process of the first two sources, several phenomena and categories emerged. Created with the visual tool MAXMap in MaxQDA, Figure 4 visualizes the emerging categories: *technology* (blue), *society* (yellow), *ethics* (red), *education* (green), and undefined categories (black). By observing the keywords in Figure 4, the alignment between keywords and categories served as the basis for the category-building process. The emerging categories emerged from the most frequently used codes identified in the documents.

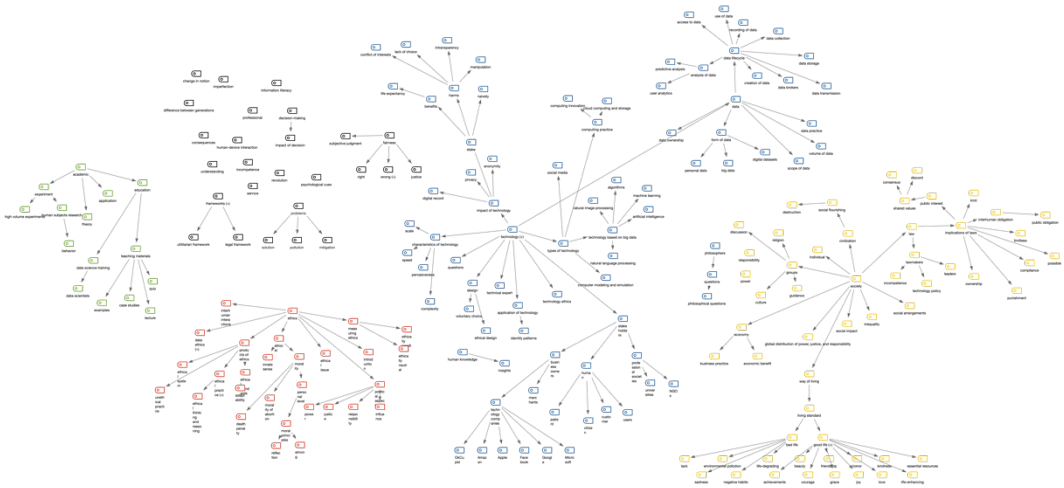

Figure 4: MAXMap after open coding the first two sources

It was also possible to define subcategories that expand the characteristics and dimensions of their parent category. For example, on the second level of *technology*, some subcategories such as *impact of technology*, *types of technology*, *characteristics of technology*, and *data* even had deeper levels, which further refined the phenomena. On the other hand, other second-level subcategories of *technology* were not yet well-developed, such as *technical experts*, *design*, *application of technology*, and *technology ethics*.

There are shared dimensions and properties between *society*, *technology*, *ethics*, and *education*. *Human subjects research*, a subcategory of *education*, concerns human behaviors, which are interhuman interactions occurring within a society, or interaction between humans and technology. Similarly, *law*, a subcategory of *society*, implies lawmakers' competence to draft and enforce technology policy, which falls under the scope of *education* and *technology*. *Ethical practice*, a subcategory of *ethics*, simultaneously refers to the *application of technology*, *social impact*, and *academic*, which are respective subcategories of *technology*, *society*, and *education*. Even though these clusters are by no means the final composition, they were able to provide a firsthand look into the thematic priorities of the first two sources. From this standpoint, the analysis continued to examine the differences and similarities between these two sources and other sources.

5.2 Categories and subcategories

After the analysis of source #1–2, it became possible to define the first categories and subcategories. According to Corbin and Strauss, a category consists of events, happenings, objects, and actions/interactions that were found to be conceptually

similar or related in meaning (1998). A category can be expanded in depth and dimensions, and similar concepts that make up a property or dimension of a category form a corresponding subcategory. To clarify the relationship between a category and its subcategories, the analyst could utilize axial coding. This process is essential for the analysis, because in this phase, not only are concepts and phenomena identified, but they will also be examined more closely in terms of their nature. In particular, several techniques could aid in this process, namely theoretical sampling, paradigm coding, or process-oriented analysis.

The analysis continued with source #3 (Curriculum framework for undergraduate and graduate students in science, health sciences, and engineering programs), #4 (Instructor Guide to Flipped Data Management Course), and #5 (Data Management Expert Guide). Figure 5 and 6 show that the most frequently identified codes in these sources are different from the first two sources:

Figure 5: Code clouds from sources #1–2

Figure 6: Code clouds from sources #3–5

In this context, commonly mentioned phenomena were *data storage/archiving/preservation*, *data sharing/transmission*, *copyright protections*, *access to data*, *research data*, *personal data*, *use and reuse of data*, and *informed consent*, which are predominantly subcategories of *technology*. Hence, with source #3, #4, and #5 being guidelines on data management for scientific research, the specializing nature was reflected clearly in the codes (Figure 5B). Subsequently, the category *technology* was expanded in depth and dimensions, especially within the subcategory *data* and *data management.*

The analyst saw the need to question the relationship between data storage/archiving/preservation, documentation of data, and data backup. A microanalysis of sentences containing these terms was an attempt to seek clarification. Whereas the criteria for storage medium of tangible objects could be dimensions, capacity, or material, data is intangible; thus, the types of media used for storage are unique to data characteristics. Due to the intangible nature of data, it is necessary to consider the longevity of data storage media. If a medium is outdated, it becomes impossible to access the data that it stores.

Therefore, when discussing data storage, it is inevitable to mention data documentation, data backup, and data migration. Data documentation describes all activities taken concerning the particular dataset and the handlings of metadata; documentation is the reference point for others to understand the handing of a dataset. Data backup ensures data storage on multiple media so that it is possible to retrieve the state of a dataset at a given point. A backup serves as a substitute or support. As with data migration, it means moving data to more viable platforms or standards when the medium in use poses a risk to become obsolete. In the end, data storage is to ensure that the data is retrievable for long-term access (archiving) and stays in compliant and stable formats (preservation), which ensure interoperability for data sharing. With other data-related measures such as documentation of data, data backup, and data protection, the process of data management started conceptualizing.

To saturate the underdeveloped subcategory *competences* under *education,* the analyst applied theoretical sampling between guidelines for natural persons (sources #1–5) and a guideline for private organizations (source #6). The Singaporean PDPC enacted the Personal Data Protection Act (PDPA) in 2012. In 2017, the PDPC published a guide for organizations on developing a data protection management program. Later in 2019, a revised version was published.

According to this guide, private Singaporean organizations should include personal data protection policies into their risk management framework. They could either appoint a Data Protection Officer (DPO) who oversees all personal data protection-related matters, a Data Protection Office with multiple positions, or outsource this position. The DPO has many responsibilities, such as ensuring compliance through data protection policies, fostering and communicating a culture for data protection, liaising with public authorities, and managing queries and complaints regarding personal data protection. Within an organization, training on personal data protection accompanies the employment journey through every step with varying levels of specialization depending on each employee's roles and responsibilities. However, for all staff, there is an onboarding briefing to communicate the fundamentals of the Singaporean PDPA and an interview to enforce requirements on proper handling of personal data upon exiting the organization. During the employment period, the organizations must periodically refresh and retrain their staff on data protection policies and processes.

Using this example, the analyst recognized similarities between data management measures of a natural person and an organization. Whereas an organization needs to appoint a DPO, each person could be their own data protection officer. However, there are also dissimilarities between an organization and a natural person regarding legal obligations. An organization is obliged to clear communication and compliant execution of personal data protection policies and processes. It will undergo regular vetting from public authorities, in this case, the PDPC. The same obligation doesn't apply to a natural person. As the PDPA applies to private organizations that collect personal data, a natural person is on the other side. Nevertheless, the PDPC recommends each individual to take responsibility for their personal data. By observing the legal obligations of private organizations, it was possible to conclude similar competencies for individuals, such as performing a routine assessment of protective measures and active initiation of queries or complaints regarding personal data.

As the analysis progressed, the sources #7-9 helped densify the existing categories. The analysis also revealed some recurring concepts–two concepts derived directly from the sources (vivo concepts), such as *informed consent* and *policy*. *Informed consent* is essential for human subjects research. When data, especially (sensitive) personal data is collected, stored, and used, the data owner must be informed on the collecting methods, potential harms, purpose of the research, and future archiving and sharing of their data. As the data owners

become aware of measures taken to protect their anonymity and the confidentiality of their data, they have the right to consent and withdraw consent at any time. Besides the content-related aspect, data owners should also receive comprehensive information on the formality of informed consent. Consent is either written, verbal, one-off or processual, granular or general; furthermore, it is also possible to retrieve consent retrospectively.

There are *policies* which decree, among other matters, how and when informed consent is needed for data sharing. The etymology of *policy* dates back to the late 14th century, from Old French policie "political organization, civil administration," from Late Latin politia "the state, civil administration," and from Greek politeia "state, administration, government, citizenship." Nowadays, the term "policy" suggests not only authorities but also organizations or institutions that exercise a certain amount of power over a body of people. It depicts the democratic aspect of policymaking. The topic policy links to policymakers, legal obligations, public interest, the relationship between ethics and law, and the relationship between law and oversight. In the context of data ethics, local institutions such as governments, universities, and research agencies often issue policies for data sharing and monitoring access to safeguard privacy, confidentiality, to establish ethical conduct, and to avoid conflicts of interests.

Furthermore, it would be inaccurate to assume that policy is permanent because even on the small scale of this analysis, the notions of policy have been irresolute. The point of policymaking is to act on problems that emerge and change continuously in scope and nature. The microanalysis on *informed consent* and *policy* amplifies the ambiguous landscape of data ethics. Concerning data ethics, context remains an indisputable factor in every consideration.

Besides vivo concepts, the analysis noticed two processes while overviewing the subcategories of *data* and *education*. First, *data management* indicates the process of managing data, from collection to sharing and preserving data. By recognizing this process, it was clear how the different steps in handling data described in the sources together portray an example of ethical practice. Second, *teaching goals* suggest the process of teaching, from the act goal setting to assessment. Teaching goals describe competencies missing or lacking before education and, at the same time, desired competencies afterward. This process assisted the analysis in identifying how the competences transform according to contextual changes in factors.

Gradually, the analysis resulted in the main categories gaining in subcategories; thus, the concept, idea, action, or phenomena behind each category became more evident thanks to the illustration by its subcategories. In terms of clarity, through the identification of vivo concepts, paradigm coding, theoretical sampling, and process-oriented analysis, it became possible to identify the nature of the relationships between subcategories, e.g., causal conditions, attributes of the context, consequences of actions, or intervening conditions (Bryant & Charmaz 2019). As the main categories started to become well-developed, the next step was to weave together the relationships between them.

5.3 The relationships between categories

At this point, it was necessary to re-analyze all eight sources to identify similarities and differences among the sources and the existing main categories. These are the categories that resulted from the inductive analysis using the grounded theory method:

Table 2: Overview of the categories from the inductive analysis

Education	Ethics	Society	Technology
academic	data ethics	societal groups	application of technology
competencies	research ethics	life quality	characteristics of technology
assessment	ethical issues	public interest	data
data science training	ethicist	economy	design of technology
human knowledge	measuring ethics	law	stakeholders
open science	political aspect	geographical boundaries	types of technology
teaching goals	products of ethics		
teaching materials	ethical review		
	codes of ethics		

Before deciding on a central category, the analyst integrated the relationships between all four categories and emphasized categories and second-level subcategories in italics:

Education* and *ethics

Many second-level subcategories of *education* intersect with dimensions of *ethics*. These are *open science, data science training,* and *academic*, whereas *academic* implies the teaching of ethics in higher education. The teaching consists of assessment via quizzes, *teaching materials* such as case studies about ethical practice or unethical practice, activities such as reviewing, reading, investigating, and *teaching goals* such as awareness about *ethical issues*. Across all sources, the most notable conjunction between *education* and *ethics* in *research ethics.* Source #5 explicitly laid out ethical skillsets for researchers while collecting, archiving, and sharing (sensitive) personal data for human subjects researches.

Education* and *technology

As a central part of scientific research, the quality of research data relies on advancing *technology* in data collection, data analysis, data storage/archiving/preservation, and the use and reuse of data. The sources indicated that research grant recipients nowadays are expected to submit a data management plan to accompany their proposals, regardless of their disciplines. Research data categorizes as a type of data, whose parent category *data* comprises other academic-relevant concepts such as data management and data sources. Among others, technical experts who undergo *academic* or vocational training, as well as universities and research agencies, are *stakeholders* in *technology* whose opinions, practices, and findings influence technology outcomes.

Ethics* and *society

For an action to be ethical, it must align with the shared values within a *society*. As stated in source #1, *ethics* is a human topic whose notions interwoven with the way of living of people in a *society*. Moreover, ethics precede *law*; each lawmaker should utilize his/her innate sense of morality to consider the potential ethical implications of legislation and make sure that the legal provisions live up to ethical standards (DEK 2019). The impacts of ethical considerations reflect the *life quality* among different *societal groups* in different areas of the world (*geographical boundaries*). The concept of legal ownership intersects with data ownership, with the latter dictating the legality of accessing, using, and sharing

data. Human civilization flourishes or decays, depending on the ethical values of its *society*. Thus, it emphasizes that ethics merely reflect shared values that the majority of *society* agree with and are embodied by *codes of ethics*.

Ethics and *technology*

Data ethics combine *ethics* and *technology*. Behind every *technology* that is beneficial to humans, there exists a conscious ethical decision. Figure 6 shows that *ethics* and *technology* go hand in hand in all sources; the content of segments coded with ethics-related codes covers an essential aspect of technology–*data*. *Technology* reflects the ethical practice of its creators, and *ethics* also plays a role in shaping technologies that serve humans and not the other way around. Through *ethical reviews* and *codes of ethics*, ethical concepts such as consent and protection find themselves becoming increasingly indispensable topics in debates about *technology*.

Figure 7: Word cloud of segments coded with ethics-related codes

Society and *technology*

On the one hand, the speed, scale, and pervasiveness of technological development in *society* are rapidly gaining in prevalence. The *stakeholders* in *technologies* represent *societal groups* such as research funders, business owners, research institutions, oversight committees, and professional societies. The social impact of technology can be measured by the *life quality* of people in different world regions. Furthermore, as *technology* is an enabler for social flourishing, legal implications apply to maintain the superior role of humans over *technology*,

especially artificial intelligence. On the other hand, *society* gives meaning to *technology* because the *application of technology* takes place in societal contexts. For instance, in an economy, users signify a problem or need, *technology* then tries to solve that problem or fulfill that need.

Education* and *society

Society is the backdrop for educational activities. In particular, social sciences researchers are tasked with synthesizing *human knowledge* and understanding social arrangements and social phenomena. Educators could be considered a *group* of professionals in *society* that play a part in forming shared values and holding a stake in social consensus or social discord. Under some circumstances, when scientific research calls for personal data collection or reveals an intimate understanding of humans, there are legal restrictions that safeguard human rights. *Laws* are established upon agreement among different societal groups to maintain a state of equilibrium between interests and benefits.
Subsequently, *education* ensures that not only people who work with data are aware of the possible consequences of their work on other people and *society*, but also lawmakers; in other words, through *education* contributes to the tending of *public interest*.

5.4 The central category

Throughout the integration of categories, the reasonings repeatedly pertain to a central theme–*stakeholders' ethical role in education, society, and technology*. "Stakeholder" is "a person such as an employee, customer, or citizen who is involved with an organization, society, etc. and therefore has responsibilities towards it and an interest in its success" (Dictionary 2020). According to Strauss requirements, the central category must be related to as many other categories and their properties as possible; the indicators pointing to the central category frequently appear in the data; it can be easily related to other data. The central category opens the doors to building the maximum variation in terms such as dimensions, properties, conditions, consequences, and strategies.

As the majority of syllabi and curricula do not require prior knowledge about data ethics, the topic is deemed suitable for stakeholders of all levels and disciplines. At the most elementary level, the training materials require that a person can name the stakeholders involved in a given situation. In the case of data ownership, the New England Collaborative Data Management Curriculum identified data as

assets, and multiple stakeholders such as institutions, funders, and scientists would want to claim their rights on this kind of intellectual property. Alternatively, in the case of data management, researchers have long been conducting their research within the legal and ethical frameworks and complying with other stakeholders' policies, e.g., sponsors, government agencies, and institutions.

The next skill that the materials proposed is the ability to distinguish and rank competing stakeholder interests. Not only is there the involvement of different stakeholders, but the guidelines also made clear that these stakeholders may not share the same interest. A person in a neutral position should be able to specify what each stakeholder wants and whose need is the priority. Source #2 made an example with a situation where there is the need to clean up a dirty dataset that will be used to train a smart pacemaker; in that instance, the primary stakeholders should be the patients who will receive the pacemaker implantation. Putting any other stakeholder's interest over the patients' would be betting against an ethically significant stake (Vallor 2018).

Even if a person is not in a neutral position and also has a stake in a data subject, the training materials suggested a mental exercise in the form of "moral imagination." This practice could help overcome mental hurdles that prevent people from doing what they know is right. The skill consists of asking oneself questions about the impact of one's choice, then going a step further to imagine the person who bears the consequences of one's action to be a friend or relative; similarly, this person could experience the joy that an ethical choice might bring. Imaginative empathy helps bridge the gap between being aware of the ethical obligations and fulfilling them.

Concerning ethics in practical application, different groups of stakeholders need different types of training due to the difference in social and educational backgrounds. To illustrate, source #6 listed two groups of stakeholders within an organization–internal and external stakeholders–and outlined the recommended measures for onboarding these stakeholders when implementing a personal data protection management program (PDPC 2019). Typically, a board of directors or senior management would need a different kind of training than general staff training; experts such as the DPO would need a completely different set of guidelines. Source #9 acknowledged the knowledge gap between stakeholders. Stakeholders can only take advantage of enhanced access to data if they possess sufficient data-awareness and skills. The knowledge gap resulted in a

disproportion between stakeholder groups, where some stakeholders who have large reserves of data and better infrastructures may also have better access to data (DEK 2019).

As backed up by the data, the category *stakeholders* have linkage to every other category. Stakeholders in education comprise roles such as educators, researchers, students. Stakeholders in society are every individual, including people with high social status such as policymakers, judges, and politicians whose actions have a direct impact on humankind. Stakeholders in technology are not necessarily individuals such as users or developers; they could also be professional societies, philanthropic organizations, and corporations. Regardless of their fields, each stakeholder acts on their moral principles; each possesses the means to become an ethicist. With *stakeholders* as the core category, it shows that current educational approaches for data ethics in higher education emphasize the awareness of learners and their active role in shaping the world regardless of their disciplines.

6 Discussion

The beginning phase of this research posed multiple questions regarding data ethics and educational approaches: How can data ethics be taught in higher education? What would be the most effective approaches? What does "data ethics" mean for different audiences, especially for educators and students? Why is it necessary to teach data ethics in higher education? How do teachings in data ethics differ pedagogically, geographically, and in different disciplines?

Regarding these questions, the findings suggest that current training approaches on data ethics in higher education center around the learner's role as a stakeholder with ethical obligations. In other words, the training materials put focus on *the ethical role of stakeholders in education, society, and technology*. All materials under analysis specified concrete actions for different stakeholders–learners, researchers, managers. According to Bloom's taxonomy (1956, for illustration see Appendices), the proposed skillsets match the levels of learning objectives:

1. Knowledge (name the stakeholders, recognize their interest).
2. Comprehension (distinguish different interests, defend the most ethically significant choice).
3. Application (fulfill ethical obligations in practical application).
4. Analyze (compare the knowledge base of different stakeholders).
5. Synthesis (combine different interests, design a data management plan).
6. Evaluation (consider potential consequences, evaluate the impact of choices).

Thus, it is an indicator, more traditional concepts could still complement educational approaches in the rather new fields of data ethics and data literacy.

In retrospect, an analysis using the grounded theory methodology was a tedious but worthwhile process. The analysis of the first two sources generated a significant amount of codes. However, at such an early stage, there was not enough data to build clusters. The codes indicated many possibilities for the way forward of the analysis. In fact, at one point, there was a significant imbalance between the number of codes for "education" and "technology" with the former being scarcely developed and the latter having multiple branches. As the process continued, the inductive approach proved useful in ensuring that the findings are grounded in data. The relationships between recurring concepts became clear. The more the progress advanced, the closer the findings inched toward theoretical

saturation, while new sources continue to add depth and dimensions to existing categories.

As Bryant reported about researches conducted by his students using the grounded theory methodology (2017), in this case, the final results did not wholly match the initial expectation of the analyst either. Even though the analyst began coding with no preconceptions, based on the literature review on data ethics and the hope to conceptualize the frameworks of teaching data ethics in higher education, it was expected that the final theory would emerge from the more technical input. Despite being resided under the primary category *technology*, *stakeholders* suggest a rather societal aspect of data ethics and apply to a broader cadre than higher education. In addition, due to the inductive nature of the approach, the analyst must decide when the categories have reached theoretical saturation, which proved to be quite challenging. Nevertheless, the amount of gathered codes was enough to reach solid conclusions.

Despite a conscious effort to achieve a diverse selection of sources, there are still geographical boundaries, as reflected in the selection of materials for analysis and worldwide discussion about ethics. Even though the world has seen a rapid increase in high-skilled data practitioners in Asian countries (Evans Data Corp 2019), the majority of scientific papers concerning data ethics training originate from the US and European countries. The research is based on materials originating predominantly from European countries and the US, except a source from Singapore. This problem echoed the disparities of ethical debates in general and data protection standards in particular in different world regions. Due to US and Europe-centric sources, topics tend to concern citizens' rights in their respective countries only. In this regard, the findings in this thesis are in no way representative of all educational approaches worldwide. Nevertheless, they offer an empirically-grounded insight into how the thematization of data ethics in higher education. The initial goal was to determine practical educational approaches for data ethics in higher education. Later on, this goal proved too advanced for this field due to state of the art. The researcher decided to narrow the goal down to examining educational approaches. Based on this research, further researches can be conducted to measure the effectiveness of the approaches under analysis.

The findings wholly align with previous insights from the literature review. First of all, they raise the topic of ethical review, ethical considerations, and

stakeholders' ethical practice. All three terms cover the preparatory measures taken to guarantee ethical standards. Both an individual and an organization can perform an ethical review. In most cases, an ethical review utilizes a Code of Ethics as the framework for evaluation. Ethical consideration is a more abstract term for ethical review. Ethical considerations denote ethical thinking and reasoning that add up to ethical practices. As the sources have reported, ethical review, ethical considerations, and ethical practice are needed as early as but not restricted to the first stages of research. An ethical review implies compliance with legal requirements, whereas ethical considerations show awareness regarding ethical practice, which then should be exercised in regular intervals to become a "by default" mentality.

Besides pointing out the premises of ethical practice, the findings suggest that there are different states of standards influencing practice, which explains the different ways stakeholders treat data. There are standards for data formats, professional standards (Code of Conduct or Code of Ethics), or legal standards. It is the responsibility of data practitioners such as researchers and system administrators to develop data management plans to ensure the protection of personal data, transparency in the process, and the interoperability of their data for sharing. For one thing, standardized data formats are essential for the reproducibility of research findings. In the case of professional standards, this is an overlooked topic. While data professionals are expected to act ethically, there are only a handful of official guidelines issued by governmental bodies such as the Code of Ethics adopted by the Croatian Committee for Ethics in Science and Higher Education, the Ethical Framework for Research approved by the Czech government, policy-relevant guidance on research ethics published by the German Ethics Council (Deutscher Ethikrat), or the Swedish Act concerning the Ethical Review of Research Involving Humans (2006:460). These guidelines predominantly target researchers and research activities, not data practices in general, for example, in software engineering or artificial intelligence.

Besides, throughout the analysis, it was noted that the sources put great focus on data management, copyright protection, and privacy, which are legitimate concerns. However, nowadays, open access, open data, and open science are gaining popularity in social discourse. As researches rely on data, more findable, accessible, interoperable, and reusable data would foster a better research culture. There need to be more discussions on the gap between data protection and the right to access information. It would be beneficial for all stakeholders if there is a

clear consensus on the impact of open data on data ownership. The sources show that concrete actions help minimize the risk of personal data exploitation and maintain data integrity, such as informed consent, anonymization, and pseudonymization. More educational guidelines on open data could serve as a companion to data protection regulations. According to the H2020 Program Guidelines on FAIR Data from the European Research Council, data should be "as open as possible and as closed as necessary" (ERC 2017).

In the long run, the findings of this research could serve as a starting point for further scientific research on the teaching of data ethics. Teaching approaches need not be stakeholder-centric; however, as proven during the analysis, it is notable to consider the stakeholder's roles and responsibilities. A human-centric educational approach is a suitable key pillar for teaching about data ethics in a human-centric technology landscape.

7 Conclusion

The research started as a quest to assess the pedagogical characteristics of educational approaches on data ethics in higher education. With the help of an inductive analysis using the grounded theory methodology on training materials of data ethics, the research was able to conceptualize the teachings of data ethics with a statement that is grounded in data. Currently, the protruding concept of data ethics in the context of data literacy training for higher education is the ethical role of stakeholders in education, society, and technology.

In essence, modern society is experiencing a technological revolution unprecedented in speed and scale. The revolution in technology brought about both positive and negative impacts. In this context, data literacy is becoming more critical than ever before. While literacy empowers humans to capture and receive knowledge, data literacy gives them the means to protect themselves and others from unfair practices and exploitations. It is proven to be an indisputable tool for not only students of all disciplines, but for anyone who has influence or is affected by data technologies.

Current progress in teaching data literacy has laid down the groundwork for data ethics. Data ethics expand on the competences of data literacy and add to data literacy a moral element. In other words, it is necessary not only to possess competencies in handling data but also to understand the potential consequences of data practice in the societal, academic, and technological contexts. There are countries, institutions, professional societies, and companies who are actively contributing to a more data-ethical world. Nonetheless, all these efforts must inevitably keep up with technological advances. Despite new challenges, stakeholders should not lose sight of the positive changes that they could make.

Although the research findings could shed light on the universal concept of current educational approaches, they have not yet been able to reflect the impact of these educational approaches. As discussed in Chapter 6, measuring the impact of training was initially a research goal. However, due to the ambiguity of data ethics as a topic, the research then prioritized first and foremost a conceptualization of educational approaches. Besides, despite the considerations concerning geographical diversity in source selection for the analysis, it was not possible to guarantee balance in the origins of materials.

Nevertheless, the research findings were able to substantiate how data ethics is being taught in higher education. Even though the literature review revealed various conflicting viewpoints regarding data ethics, in the end, the findings showed that the roles and responsibilities of stakeholders remain a quintessential factor. The findings could serve as a reference point for future training programs. On the grounds of a stakeholder-centric approach, future training curricula, competence frameworks, and future researches on teaching data ethics could benefit from a unified understanding of pedagogical goals in higher education. Data professionals should receive a thorough education on the Code of Ethics within their professions. More importantly, training programs could elevate students' self-awareness as future experts in their field, regardless of their disciplines. For example, law students would later take part in policymaking and need a basic understanding of data ethics to judge policies fairly; students in political sciences need data ethics to assess data practice in politics more critically. Ethical data competencies are needed in all walks of life, and the human factor decides if the best scenario could happen.

It would be encouraging to see more training programs with a more inclusive view on diversity, especially in the context of globalization. In addition to diversity, the findings prepared the groundwork for research on teaching data ethics in mainstream education. Younger generations are exposed to new technologies at much younger ages than their seniors; hence, they also strongly need guidance on the right mentality towards data. Lastly, despite being a less prominent theme among the findings, open data versus data protection could be a topic for future researches, as different stakeholders may have different views on the subject. A better understanding of this topic could bring more clarity to the curricula of data ethics training.

Bibliography

ABRAHAMSON, Eric, 1996, Management Fashion. *The Academy of Management Review*. 1996. Vol. 21, no. 1, p. 254. DOI 10.2307/258636.

ACHENBACH, Joel, 2015, Techno-skeptics' objection growing louder. *Chicago Tribune* [online]. 28 December 2015. Available from: https://www.chicagotribune.com/business/blue-sky/ct-wp-technoskeptics-objection-growing-louder-bsi-20151226-story.html

ALPHABET INC., [no date], Code of Conduct [online][viewed 28 July 2020]. Available from: https://abc.xyz/investor/other/code-of-conduct.

AMNESTY INTERNATIONAL, 2019, *Surveillance Giants: How the business model of Google and Facebook threatens human rights* [online][viewed 20 July 2020]. Available from: https://www.amnesty.org/download/documents/pol3014042019english.pdf.

ANDERSON, Carl, 2015, *Creating a data-driven organization: practical advice from the trenches*. O'Reilly.

ASSOCIATION OF COLLEGE AND RESEARCH LIBRARIES, 2000, ACRL STANDARDS: Information Literacy Competency Standards for Higher Education. *College & Research Libraries News.* Vol. 61, no. 3, pp. 207-215. DOI: 10.5860/crln.61.3.207.

ACRL Information Literacy Competency Standards Review Task Force, 2012, *Task Force Recommendations* [online]. ACRL AC12 Doc 13.1, June 2, 2012 [viewed 22 July 2020]. Available from: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/standards/ils_recomm.pdf.

ASSOCIATION OF COLLEGE AND RESEARCH LIBRARIES, 2015, Framework for Information Literacy for Higher Education [online]. American Library Association [viewed 22 July 2020]. Available from: http://www.ala.org/acrl/files/issues/infolit/framework.pdf.

ATKINS, Daniel E., BROWN, John Seely and HAMMOND, Allen L., 2007, *A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities* [online]. Report to the William and Flora Hewlett Foundation [viewed 22 July 2020]. Available from: https://hewlett.org/wp-content/uploads/2016/08/ReviewoftheOERMovement.pdf.

BARRAT, James, 2015, *Our final invention: artificial intelligence and the end of the human era*. Thomas Dunne Books.

BEAUMONT, Peter, 2013, NSA leaks: US and Britain team up on mass surveillance. *The Guardian* [online]. 22 June 2013 [viewed 20 July 2020]. Available from: https://www.theguardian.com/world/2013/jun/22/nsa-leaks-britain-us-surveillance.

BEIER, Michael, 2018, Digitale Strategien für Nonprofit-Organisationen Anfang des 21. Jahrhunderts (Digital Strategies for Nonprofit Organizations at the Beginning of the 21st Century). *SSRN Electronic Journal*. 2018. DOI 10.2139/ssrn.3316052.

BIK, Holly M. and GOLDSTEIN, Miriam C., 2013, An Introduction to Social Media for Scientists. *PLoS Biology*. 2013. Vol. 11, no. 4. DOI 10.1371/journal.pbio.1001535.

BLOOM, Benjamin S., ENGELHART, Max D., FURST, Edward J., HILL, Walker H. and KRATHWOHL, David R., 1956, *Taxonomy of educational objectives: the classification of educational goals*. Longman.

BOONE, Brian, 2017, *ETHICS 101: From altruism and utilitarianism to bioethics and political ethics, an exploration of the concepts of right and wrong*. ADAMS MEDIA Corporation. ISBN 9781507204948 1507204949.

BOYD, Danah, 2005, *Why Web2.0 Matters: Preparing for Glocalization* [online]. Zephoria.org [viewed 18 July 2020]. Available from: http://www.zephoria.org/thoughts/archives/2005/09/05/why_web20_matte.html.

BRYANT, Antony, 2017, *Grounded theory and grounded theorizing: pragmatism in research practice.* Oxford University Press.

BRYANT, Anthony J. and CHARMAZ, Kathy, 2019, *The SAGE Handbook of Current Developments in Grounded Theory*. SAGE Reference.

BRYNJOLFSSON, Erik, HITT, Lorin M. and KIM, Heekyung Hellen, 2011, Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance? *SSRN Electronic Journal*. 2011. DOI 10.2139/ssrn.1819486.

CARLSON, Jake and BRACKE, Marianne, 2015, Planting the Seeds for Data Literacy: Lessons Learned from a Student-Centered Education Program,

International Journal of Digital Curation. 2015. Vol. 10, no. 1. DOI 10.5703/1288284315518.

CESSDA Training Team (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. DOI: 10.5281/zenodo.3820473

CHIGNARD, Simon, 2013, *A brief history of Open Data* [online][viewed 28 July 2020], Available from: http://www.paristechreview.com/2013/03/29/brief-history-open-data.

CLARKE, Adele E., 2019, Situating Grounded Theory and Situational Analysis in Interpretive Qualitative Inquiry. *The SAGE Handbook of Current Developments in Grounded* Theory. 2019. Pp 3–48. DOI 10.4135/9781526485656.

CRUSOE, David, 2016, Data Literacy defined pro populo: To read this article, please provide a little information [online]. *The Journal of Community Informatics*, vol. 12, no. 3 [viewed 18 May 2020]. Available from http://www.ci-journal.net/index.php/ciej/article/view/1290.

CUNNINGHAM, R. L., 1967, Ethics and game theory: The prisoner's dilemma. Public *Choice*. 1967. Vol. 2, no. 1, p. 11–26. DOI 10.1007/bf01718649.

DATA ETHICS COMMISSION OF THE FEDERAL GOVERNMENT, 2019, *Opinion of the Data Ethics Commission: Executive Summary* [online][viewed 16 August 2020]. Available from: https://datenethikkommission.de/wp-content/uploads/191023_DEK_Kurzfassung_en_bf.pdf.

DAVIES, Harry, 2015, Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users. *The Guardian* [online]. 11 December 2015 [viewed 20 July 2020]. Available from: https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

DIEBOLD, Francis X., 2012, On the Origin(s) and Development of the Term 'Big Data'. *SSRN Electronic Journal*. 2012. DOI 10.2139/ssrn.2152421.

DUNLAP, Karen and PIRO, Jody S., 2016, Diving into data: Developing the capacity for data literacy in teacher education. *Cogent Education*. 2016. Vol. 3, no. 1. DOI 10.1080/2331186x.2015.1132526.

EDINA AND DATA LIBRARY, UNIVERSITY OF EDINBURGH, 2017, *Research Data MANTRA* [online]. Online training units & data handling tutorials

[viewed 22 July 2020]. Zenodo. Available from: http://doi.org/10.5281/zenodo.1035218.

EMROUZNEJAD, Ali and CHARLES, Vincent, 2019, *Big data for the greater good*. Springer.

EUROPEAN RESEARCH COUNCIL, 2017. *Guidelines on Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council under Horizon 2020* [online]. 21 April 2017 [viewed 16 August 2020]. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf.

EVANS DATA CORP, 2019, *Global Developer Population and Demographic Study 2019*, vol. 1. Santa Cruz: Evans Data Corporation.

FLORIDI, Luciano and TADDEO, Mariarosaria, 2016, What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016. Vol. 374, no. 2083. DOI 10.1098/rsta.2016.0360.

FREILAND, Dirk, 2016, The worldwide consulting industry: Analysis 2016 - focus market and trends in Germany [online]. *Clairfield International consulting industry market study* [viewed 19 July 2020]. Available from: https://www.slideshare.net/dirkfreiland/clairfield-international-consulting-industry-market-study-part-1.

FUCHS, Christian, 2008, *Internet and society: social theory in the Internet age*. Pp 125-127. Routledge.

GABRIEL, Deborah, 2020, Inductive and deductive approaches to research. *Dr. Deborah Gabriel* [online]. 13 April 2020 [viewed 18 July 2020]. Available from: https://deborahgabriel.com/2013/03/17/inductive-and-deductive-approaches-to-research.

GHOSH, Bhaskar, BURDEN, Adam, and WILSON, James, 2019, *How to scale innovation and achieve full value with Future Systems* [online]. Accenture [viewed 19 July 2020]. Available from: https://www.accenture.com/_acnmedia//thought-leadership-assets/pdf/accenture-future-systems-report.pdf.

GIBSON, J. Phil and MOURAD, Teresa, 2018, The growing importance of data literacy in life science education. *American Journal of Botany*. 2018. Vol. 105, no. 12, pp. 1953–1956. DOI 10.1002/ajb2.1195.

GOTZ, David and BORLAND, David, 2016, Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. *IEEE Computer Graphics and Applications*. 2016. Vol. 36, no. 3, pp. 90–96. DOI 10.1109/mcg.2016.59.

GRAHAM-HARRISON, Emma and CADWALLADR, Carole, 2018, Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian* [online]. 17 March 2018 [viewed 1 June 2020]. Available from: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

GRASSEGGER, Hannes and KROGERUS, Mikael, 2016, Ich habe nur gezeigt, dass es die Bombe gibt. *Das Magazin* [online]. 8 December 2016 [viewed 20 July 2020]. Available from: https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt.

GREENWALD, Glenn and MACASKILL, Ewen, 2013, NSA Prism program taps in to user data of Apple, Google and others. *The Guardian* [online]. 7 June 2013 [viewed 20 July 2020]. Available from: https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data.

GRIFFIN, Jane, 2008, The Role of the Chief Data Officer. DM Review, vol. 18, no. 2, p. 28. ISSN 1521-2912.

HOLST, Arne, 2020, Data created worldwide 2010-2024. *Statista* [online]. 7 July 2020. Available from: https://www.statista.com/statistics/871513/worldwide-data-created.

HUNTER, David and EVANS, Nicholas, 2016, Facebook emotional contagion experiment controversy. *Research Ethics*. 2016. Vol. 12, no. 1, p. 2–3. DOI 10.1177/1747016115626341.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* [online], First Edition [viewed 29 July 2020]. Available from: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

INFORMATION: meaning in the Cambridge English Dictionary, [no date]. *Cambridge Dictionary* [online][viewed 17 June 2020], Available from: https://dictionary.cambridge.org/dictionary/english/information.

INFORMATION COMMISSIONER'S OFFICE, 2018, *Investigation into the use of data analytics in political campaigns* [online][viewed 20 July 2020]. Available from: https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf.

INGRAM, David, 2018, Factbox: Who is Cambridge Analytica and what did it do? *Reuters* [online]. 20 March 2018 [viewed 20 July 2020]. Available from: https://www.reuters.com/article/us-facebook-cambridge-analytica-factbox-iduskbn1gw07f.

JANOWSKI, Tomasz, 2015, Digital government evolution: From transformation to contextualization. *Government Information Quarterly*. 2015. Vol. 32, no. 3, pp. 221–236. DOI 10.1016/j.giq.2015.07.001.

JOHNSTON, Lisa R. and JEFFRYES, Jon, 2014, *Instructor Guide to the Flipped Data Management Course* [online]. University of Minnesota Libraries [viewed 22 July 2020]. Available from: http://z.umn.edu/teachdatamgmt.

JOULEVA, Gergana et al. Human rights groups' open letter to David Cameron on surveillance, 2013. *The Guardian* [online][viewed 20 July 2020]. Available from: https://www.theguardian.com/commentisfree/2013/nov/03/human-rights-groups-letter-david-cameron.

KAPLAN, Andreas M. and HAENLEIN, Michael, 2010, Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*. 2010. Vol. 53, no. 1, pp. 59–68. DOI 10.1016/j.bushor.2009.09.003.

KAPOOR, Kawaljeet Kaur, TAMILMANI, Kuttimani, RANA, Nripendra P. et al., 2018, Advances in Social Media Research: Past, Present and Future. *Inf Syst Front*, vol. 20, pp. 531–558. DOI: https://doi.org/10.1007/s10796-017-9810-y.

KATAL, Avita, WAZID, Mohammad and GOUDAR, R. H., 2013, Big data: Issues, challenges, tools and Good practices. *2013 Sixth International Conference on Contemporary Computing (IC3)*. DOI 10.1109/ic3.2013.6612229.

KEEN, Andrew, 2015, *The Internet is Not the Answer*. Atlantic Books Ltd.

KITCHIN, Rob, 2014, *The data revolution: big data, open data, data infrastructures and their consequences*. Sage.

KRAMER, Adam D. I., GUILLORY, Jamie E. and HANCOCK, Jeffrey T., 2014, Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*. 2014. Vol. 111, no. 24, pp. 8788–8790. DOI 10.1073/pnas.1320040111.

KRISHNAMURTHY, Balachander and WILLS, Craig E., 2009, On the leakage of personally identifiable information via online social networks. Proceedings of the 2nd ACM *workshop on Online social networks - WOSN '09*. 2009. DOI 10.1145/1592665.1592668.

LAMAR SOUTTER LIBRARY, UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL, [no date], *New England Collaborative Data Management Curriculum* [online][viewed 22 July 2020]. Available from: https://library.umassmed.edu/resources/necdmc/index.

LANIER, Jaron, 2011, *You are not a gadget: a manifesto*. Penguin.

LEARN2ANALYZE, 2020, Learn to Analyze Educational Data and Improve your Blended and Online Teaching [online]. *Learn2Analyze MOOC* [viewed 16 August 2020]. Available from: https://learn2analyse.eu/proj/l2a-mooc.

LEONELLI, Sabina, 2016, Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016. Vol. 374, no. 2083. DOI 10.1098/rsta.2016.0122.

LLOYD-SMITH, Lindsay, 2011, *DataTrain: Open Access Post-Graduate Teaching Materials in Managing Research Data in Archaeology* [online]. University of Cambridge [viewed 22 July 2020]. Available from: https://archaeologydataservice.ac.uk/learning/DataTrain.xhtml.

LOHR, Steve, 2012, The Age of Big Data. *The New York Times* [online]. 11 February 2012 [viewed 27 July 2020]. Available from: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

MANDUCA, Cathryn A. and MOGK, David W., 2002. *Using Data in Undergraduate Science Classrooms* [online]. Final report. Northfield: Carleton

College [viewed 18 Juy 2020]. Available from: https://serc.carleton.edu/files/usingdata/usingdata.pdf.

MARX, Vivien, 2013, The big challenges of big data. *Nature*. 2013. Vol. 498, no. 7453, pp. 255–260. DOI 10.1038/498255a.

MASON, Jeff, 2014, Obama takes swipe at Snowden in spy reform speech. *Reuters* [online]. 17 January 2014 [viewed 20 July 2020]. Available from: https://www.reuters.com/article/us-usa-security-obama-snowden-idUSBREA0G1DW20140117.

MAYBEE, Clarence and ZILINSKI, Lisa, 2015, Data informed learning: A next phase data literacy framework for higher education. *Proceedings of the Association for Information Science and Technology*. 2015. Vol. 52, no. 1, p. 1–4. DOI 10.1002/pra2.2015.1450520100108.

MAYBEE, Clarence, CARLSON, Jake, SLEBODNIK, Maribeth and CHAPMAN, Bert, 2015, "It's in the Syllabus": Identifying Information Literacy and Data Information Literacy Opportunities Using a Grounded Theory Approach. *The Journal of Academic Librarianship*. 2015. Vol. 41, no. 4, pp. 369–376. DOI 10.1016/j.acalib.2015.05.009.

MERVIS, Jeffrey, 2012, Agencies Rally to Tackle Big Data. *Science*. 2012. Vol. 336, no. 6077, p. 22. DOI 10.1126/science.336.6077.22.

MING, Wu and HUI, Hu, 2018, Data Literacy Education Design Based on Needs of Graduate Students in University of Chinese Academy of Sciences. *Communications in Computer and Information Science Information Literacy in the Workplace*. 2018. Pp. 158–168. DOI 10.1007/978-3-319-74334-9_17.

MITTELSTADT, Brent Daniel, ALLO, Patrick, TADDEO, Mariarosaria, WACHTER, Sandra and FLORIDI, Luciano, 2016, The ethics of algorithms: Mapping the debate. *Big Data & Society*. 2016. Vol. 3, no. 2. DOI 10.1177/2053951716679679.

MITTELSTADT, Brent, 2019, Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*. 2019. Vol. 1, no. 11, p. 501–507. DOI 10.1038/s42256-019-0114-4.

MITYAGIN, Sergey A., DROJJIN, Sergey I. and TIKHONOVA, Olga B., 2017, A Value-Oriented Approach in Smart City Projects Selection and Ranking.

Communications in Computer and Information Science Digital Transformation and Global Society. 2017. Pp. 307–318. DOI 10.1007/978-3-319-69784-0_26.

NEWPORT, Frank, 2013, Americans Disapprove of Government Surveillance Programs. *Gallup.com* [online]. 12 June 2013 [viewed 20 July 2020]. Available from: https://news.gallup.com/poll/163043/americans-disapprove-government-surveillance-programs.aspx.

NGAI, Eric W.t., TAO, Spencer S.c. and MOON, Karen K.l., 2015, Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*. 2015. Vol. 35, no. 1, pp. 33–44. DOI 10.1016/j.ijinfomgt.2014.09.004.

NGUYEN, Dan, 2014, *Public Affairs Data Journalism* [online]. Syllabus, assignments and other resources. Stanford Computational Journalism Lab [viewed 22 July 2020]. Available from: http://www.padjo.org/lectures/comm_273d/2014_fall/syllabus.

NOORDEN, Richard Van, 2014, Online collaboration: Scientists and the social network. *Nature*. 2014. Vol. 512, no. 7513, pp. 126–129. DOI 10.1038/512126a.

O'CONNOR, Lisa and WONG, Gabrielle K. W., 2010, Facilitating Students' Intellectual Growth in Information Literacy Teaching. *Reference & User Services Quarterly*. 2010. Vol. 50, no. 2, pp. 114–118. DOI 10.5860/rusq.50n2.114.

OPEN KNOWLEDGE FOUNDATION, 2020, *Open Definition 2.1* [online]. Open Defintion [viewed 22 July 2020]. Available from: https://opendefinition.org/od/2.1/en.

OPEN KNOWLEDGE FOUNDATION, 2020, *Training* [online][viewed 22 July 2020]. Available from: https://okfn.org/what-we-do/training.

PALFREY, John and GASSER, Urs, 2008, *Born digital: understanding the first generation of digital natives*. Basic Books.

PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2019, *Guide to developing a data protection management program* [online]. 15 July 2019 [viewed 16 August 2020]. Available from: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/DPMP/Guide-to-Developing-a-Data-Protection-Management-Programme-15-July-2019.pdf?la=en.

PEW RESEARCH CENTER, 2013, Majority Views NSA Phone Tracking as Acceptable Anti-terror Tactic. *Pew Research Center - U.S. Politics & Policy* [online]. 10 June 2013 [viewed 20 July 2020]. Available from: https://www.pewresearch.org/politics/2013/06/10/majority-views-nsa-phone-tracking-as-acceptable-anti-terror-tactic.

PRADO, Javier Calzada and MARZAL, Miguel Ángel, 2013, Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. *Libri*. 2013. Vol. 63, no. 2. DOI 10.1515/libri-2013-0010.

PRENSKY, Marc, 2001, Digital natives, digital immigrants. *On the Horizon*. 2001. Vol. 9, no. 5. MCB University Press.

QIN, Jian and D'IGNAZIO, John, 2010, Lessons learned from a two-year experience in science data literacy education [online]. *International Association of Scientific and Technological University Libraries, 31st Annual Conference*. Paper 5 [viewed 18 July 2020]. Available from: http://docs.lib.purdue.edu/iatul2010/conf/day2/5.

QUATTROCIOCCHI, Walter, SCALA, Antonio and SUNSTEIN, Cass R., 2016, Echo Chambers on Facebook. *SSRN Electronic Journal*. 2016. DOI 10.2139/ssrn.2795110.

RABATÉ Jean-Michel, 2008, *The ethics of the lie*. New York: Other Press, 2007. ISBN 9781590512692 1590512693.

RASMUSSEN_POLL, 2013, 59% Oppose Government's Secret Collecting of Phone Records. *Rasmussen Reports*® [online][viewed 20 July 2020]. Available from: https://www.rasmussenreports.com/public_content/politics/general_politics/june_2013/59_oppose_government_s_secret_co.

REFORM GOVERNMENT SURVEILLANCE, 2020. *Principles* [online][viewed 20 July 2020]. Available from: https://www.reformgovernmentsurveillance.com/principles.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119 4.5.2016, p. 1.

REIS, João, AMORIM, Marlene, MELÃO, Nuno and MATOS, Patrícia, 2018, Digital Transformation: A Literature Review and Guidelines for Future Research. *Advances in Intelligent Systems and Computing Trends and Advances in Information Systems and Technologies*. 2018. Pp. 411–421. DOI 10.1007/978-3-319-77703-0_41.

RICE, Robin, 2011, JISC Research Data MANTRA Final Report [online]. University of Edinburgh [viewed 22 July 2020]. Available from: http://www.ed.ac.uk/schools-departments/information-services/about/organisation/edl/data-library-projects/mantra/deliverables.

RIDSDALE, Chantel, ROTHWELL, James, SMIT, Mike, ALI-HASSAN, Hossam, BLIEMEL, Michael, IRVINE, Dean, KELLEY, Dan E., MATWIN, Stan S. and WUETHERICK, Brad, 2015, *Strategies and best practices for data literacy education: Knowledge synthesis report* [online]. Available from: DOI: 10.13140/RG.2.1.1922.5044. (Accessed: 17 Apr 2020)

ROSS, Jeanne W., SEBASTIAN, Ina M., BEATH, Cynthia, SCANTLEBURY, Stuart, MOCKER, Martin, FONSTAD, Nils, KAGAN, Martin, MOLONEY, Kate and GERAGHTY KRUSEL, Susan, 2016, Designing Digital Organizations. *CISR Working Paper*, no. 46. MIT Center for IS Research.

ROUSSEAU, Jean-Jacques, 1950, *The social contract, and Discourses.* New York: Dutton (Original work published in 1762).

ROYAL SOCIETY, 2019a, Data skills for all [online]. *Note of discussions at a Royal Society and STEM Learning workshop on data science skills*, 13 September 2019, National STEM Learning Centre, University of York [viewed 23 July 2020]. Available from: https://royalsociety.org/-/media/policy/projects/dynamics-of-data-science/data-skills-for-all-royal-society-STEM-learning-autumn-2019.pdf.

ROYAL SOCIETY, 2019b, Dynamics of data science skills: How can all sectors benefit from data science talent?. ISBN: 978-1-78252-395-6. Also available in PDF from: http://www.royalsociety.org/dynamics-of-data-science-skills.

SAYRE-MCCORD, Geoff, 2012, Metaethics [online]. *Stanford Encyclopedia of Philosophy* [viewed 20 August 2020]. Available from: https://plato.stanford.edu/entries/metaethics.

SCHIELD, Milo, 2004, Information Literacy, Statistical Literacy and Data Literacy [online]. *IASSIST Quarterly Summer/Fall 2004*, pp. 6-11 [viewed 1 June 2020]. Available from: DOI 10.1.1.144.6309.

SCHÜLLER, Katharina, BUSCH, Paulina, and HINDINGER, Carina, 2019. *Hochschulforum Digitalisierung Arbeitspapier 47: Future Skills: Ein Framework für Data Literacy* [online][viewed 18 July 2020]. Available from: http://doi.org/10.5281/zenodo.3349865.

SMALHEISER, Neil R., 2017, *Data literacy: how to make your experiments robust and reproducible*. Academic Press.

SNOWDEN, Edward, 2019, *Permanent Record*. Metropolitan Books.

STAKEHOLDER: meaning in the Cambridge English Dictionary, [no date]. *Cambridge Dictionary* [online][viewed 16 August 2020], Available from: https://dictionary.cambridge.org/dictionary/english/stakeholder.

STANFORD ENCYCLOPEDIA OF PHILOSOPHY, [no date], *Search query "ethics"* [online][viewed 27 July 2020]. Available from: https://plato.stanford.edu/search/searcher.py?query=ethics.

STATISTA RESEARCH DEPARTMENT, 2013, *NSA's secret data collection - public opinion* [online]. 12 June 2013 [viewed 20 July 2020]. Available from: https://www.statista.com/statistics/260140/opinion-of-americans-on-whether-the-nsas-secret-data-collection-is-acceptable.

STEPHENSON, Elizabeth and CARAVELLO, Patti Schifter, 2007, Incorporating data literacy into undergraduate information literacy programs in the social sciences. *Reference Services Review*. 2007. Vol. 35, no. 4, pp. 525–540. DOI 10.1108/00907320710838354.

STRAUSS, Anselm L., 1987, *Qualitative analysis for social scientists*. Cambridge University Press. ISBN 0 521 32845 4.

STRAUSS, Anselm L. and CORBIN, Juliet M., 1990, Grounded theory research: procedures, canons, and evaluative criteria. *Qualitative Sociology*, vol. 13, pp. 3–21. DOI 10.1007/BF00988593.

STRAUSS, Anselm L. and CORBIN, Juliet M., 1998, *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications, Thousand Oaks.

STRAUSS, Anselm L. and CORBIN, Juliet M., 2008. *Basics of Qualitative Research, 3rd Edition*. Sage Publications, Thousand Oaks.

TAYLOR, Paul, 2012, What is big data? [online]. *Financial Times* [viewed 28 July 2020]. Available from: https://www.ft.com/content/5cf5751a-4077-11e2-8f90-00144feabdc0.

TAYLOR, Astra, 2015, *The people's platform: taking back power and culture in the digital age*. Picador.

TEGMARK, Max, 2018, *Life 3.0: being human in the age of artificial intelligence*. Penguin Books.

TESLA, 2016a, All Tesla Cars Being Produced Now Have Full Self-Driving Hardware. *Tesla, Inc* [online]. 1 December 2016 [viewed 28 July 2020]. Available from: https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware.

TESLA, 2016, A Tragic Loss. *Tesla, Inc* [online]. 30 June 2016 [viewed 28 July 2020]. Available from: https://www.tesla.com/blog/tragic-loss.

THOMSON, Judith Jarvis., 1985, *The trolley problem*. Faculty of Law, University of Toronto.

TING, Yu-Liang, 2015, Tapping into students' digital literacy and designing negotiated learning to promote learner autonomy. *The Internet and Higher Education*. 2015. Vol. 26, pp. 25–32. DOI 10.1016/j.iheduc.2015.04.004.

TRANSBERG, Pernille, HASSELBALCH, Gry, OLSEN, Birgitte K. and BYRNE, Catrine S., 2018, *DATAETHICS – Principles and Guidelines for Companies, Authorities & Organisations.* AKAPRINT A/S. ISBN 9788771920475.

UBALDI, Barbara, 2013, Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris. DOI 10.1787/5k46bj4f03s7-en.

UN ECONOMIC AND SOCIAL COUNCIL, 2020, Report of the Secretary-General [online]. *Progress towards the Sustainable Development Goals* [viewed 18 August 2020]. Available from: https://sustainabledevelopment.un.org/content/documents/26158Final_SG_SDG_Progress_Report_14052020.pdf

UNESCO and COMMONWEALTH OF LEARNING, 2015, Guidelines for Open Educational Resources (OER)in Higher Education [online][viewed 22 July 2020]. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000213605.locale=en.

UN GENERAL ASSEMBLY, 2015, *Transforming our world: the 2030 Agenda for Sustainable Development* [online], A/RES/70/1 [viewed 18 July 2020]. Available from: https://www.refworld.org/docid/57b6e3e44.html.

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL, LAMAR SOUTTER LIBRARY, and WORCESTER POLYTECHNIC INSTITUTE, GEORGE C. GORDON LIBRARY, 2012, *Frameworks for a Data Management Curriculum: Course Plans for Data Management Instruction to Undergraduate and Graduate Students in Science, Health Sciences and Engineering Programs* [online]. Amherst, MA: University of Massachusetts Medical School Lamar Soutter Library [viewed 22 July 2020]. Available from: https://library.umassmed.edu/pdfs/data_management_frameworks.pdf.

UN STATISTICS DIVISION, 2019, *Countries that adopt and implement constitutional, statutory and/or policy guarantees for public access to information* [online][viewed 24 June 2020]. Available from: https://ourworldindata.org/grapher/countries-that-adopt-guarantees-for-public-access-to-information.

VALLOR, Shannon, 2018, An Introduction to Data Ethics. *Introductory ethics module for data science courses* [online], Santa Clara University, Markkula Center for Applied Ethics [viewed 24 July 2020]. Available from: https://www.scu.edu/media/ethics-center/technology-ethics/IntroToDataEthics.pdf.

WATSON, Hugh J., 2014, Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. *Communications of the Association for Information Systems*. 2014. Vol. 34, no. 65, pp. 1247–1268. DOI 10.17705/1cais.03465.

What is DataONE? [online]. DataONE, © 2020 [viewed 22 July 2020]. Available from: https://old.dataone.org/what-dataone.

Willkommen bei der Hamburg Open Online University!. Hamburg Open Online University, © 2020 [viewed 22 July 2020]. Available from: https://www.hoou.de.

WISEMAN, Jane M., 2018, *Data-Driven Government: The Role of Chief Data Officers* [online]. IBM Center for The Business of Government [viewed 28 July

2020]. Available from: http://www.businessofgovernment.org/report/data-driven-government-role-chief-data-officers.

WOLFF, Annika, GOOCH, Daniel, CAVERO MONTANER, Jose J, RASHID, Umar, and KORTUEM, Gerd, 2016, Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*. Vol. 12, no. 3, pp. 9-26. ISSN: 1721-4441.

WORLD ECONOMIC FORUM, 2020, *Digital Transformation: Powering the Great Reset* [online]. Community Paper [viewed 19 July 2020]. Available from: http://www3.weforum.org/docs/WEF_Digital_Transformation_Powering_the_Great_Reset_2020.pdf.

ZIKOPOULOS, Paul, 2012, *Understanding big data: analytics for enterprise class Hadoop and streaming data*. McGraw-Hill.

ZIMMER, Michael, 2010, "But the data is already public": on the ethics of research in Facebook. *The Ethics of Information Technologies*, vol. 12, pp. 313-325. DOI 10.1007/s10676-010-9227-5.

ZIMMER, Michael, 2018, Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society*. 2018. Vol. 4, no. 2, pp. 1-11. DOI 10.1177/2056305118768300.

Appendices

Materials on data literacy training with regard to data ethics

Nr.	Name	Provider/Country	Year	License for re-use
1	Data Science Ethics	University of Michigan (USA)	2020	All rights reserved
2	An Introduction to Data Ethics	Markkula Center for Applied Ethics at Santa Clara University (USA)	2018	Free with permission to use
3	Curriculum framework for undergraduate and graduate students in science, health sciences, and engineering programs (Module 4, 5, 6)	Lamar Soutter Library, University of Massachusetts Medical School and the George C. Gordon Library, Worcester Polytechnic Institute (USA)	2012	CC BY-NC-SA 3.0
4	Instructor Guide to Flipped Data Management Course	University of Minnesota (USA)	2014	CC BY-NC 4.0
5	Data Management Expert Guide (Introduction & Chapter 5)	CESSDA Training Working Group (Consortium of European Social Science Data Archives) (EU)	2017–2019	CC BY-SA 4.0
6	Guide to Developing a Data Protection Management Program	Personal Data Protection Commission (Singapore)	2019	All rights reserved

7.1	New England Collaborative Data Management Curriculum (Module 1)	UMass Medical School, Lamar Soutter Library (USA)	Since 2012	CC BY-NC-SA 4.0
7.2	New England Collaborative Data Management Curriculum (Module 5)	UMass Medical School, Lamar Soutter Library (USA)	Since 2012	CC BY-NC-SA 4.0
8	Learn to Analyze Educational Data and Improve your Blended and Online Teaching Massive Open Online Course (MOOC)	Learn2Analyze — An Academia-Industry Knowledge Alliance for enhancing Online Training Professionals' (Instructional Designers and e-Trainers) (EU)	2020	All rights reserved
9	Opinion of the Data Ethics Commission (Executive Summary)	Data Ethics Commission (Germany)	2019	All rights reserved
10	Data Analyst in Python	Dataquest Labs, Inc.	2020	All rights reserved.
11	The Data Literacy Project	QlikTech International AB	2020	All rights reserved
12	Data Storytelling	QlikTech International AB	2020	All rights reserved
13	A Culture of Data Literacy	QlikTech International AB	2020	All rights reserved
14	Data-informed Decision	QlikTech		

	Making	International AB		
15	Data Training for Professionals	StackFuel GmbH	2020	All rights reserved
16	Auffinden, Zitieren, Dokumentieren	ZBW, GESIS, RatSWD	2020	All rights reserved
17	Data Playbook Toolkit	Heather Leson (IFRC), Dirk Slater (Fabriders)	2020	CC BY 3.0
18	Visualization for Data Journalism	University of Illinois	2020	All rights reserved
19	What is data? What is data literacy?	Eastern Michigan University Library	2018	CC BY-SA 4.0
20	Data-driven Decision Making	pwc	2020	All rights reserved
21	Humane Technology Design Guide	Center for Humane Technology	2019	All rights reserved
22	Academic Information Seeking	University of Copenhagen; Technical University of Denmark (DTU)	2020	All rights reserved
23	Being a researcher (in Information Science and Technology)	Politecnico di Milano	2020	All rights reserved
24	Information Literacy Standards for Higher Education	ACRL	2015	CC BY-NC-SA 4.0
25	Data Ethics Framework	UK Department for Digital, Culture, Media & Sport	2020	Open Government Licence v3.0

26	Personal Data Protection Act E-Learning Program	Personal Data Protection Commission (Singapore)	2018	All rights reserved
27	People+AI Guidebook	Google PAIR	2019	CC BY-NC-SA 4.0
28	Ethics, Technology and Engineering	Eindhoven University of Technology	2020	All rights reserved

Bloom's taxonomy

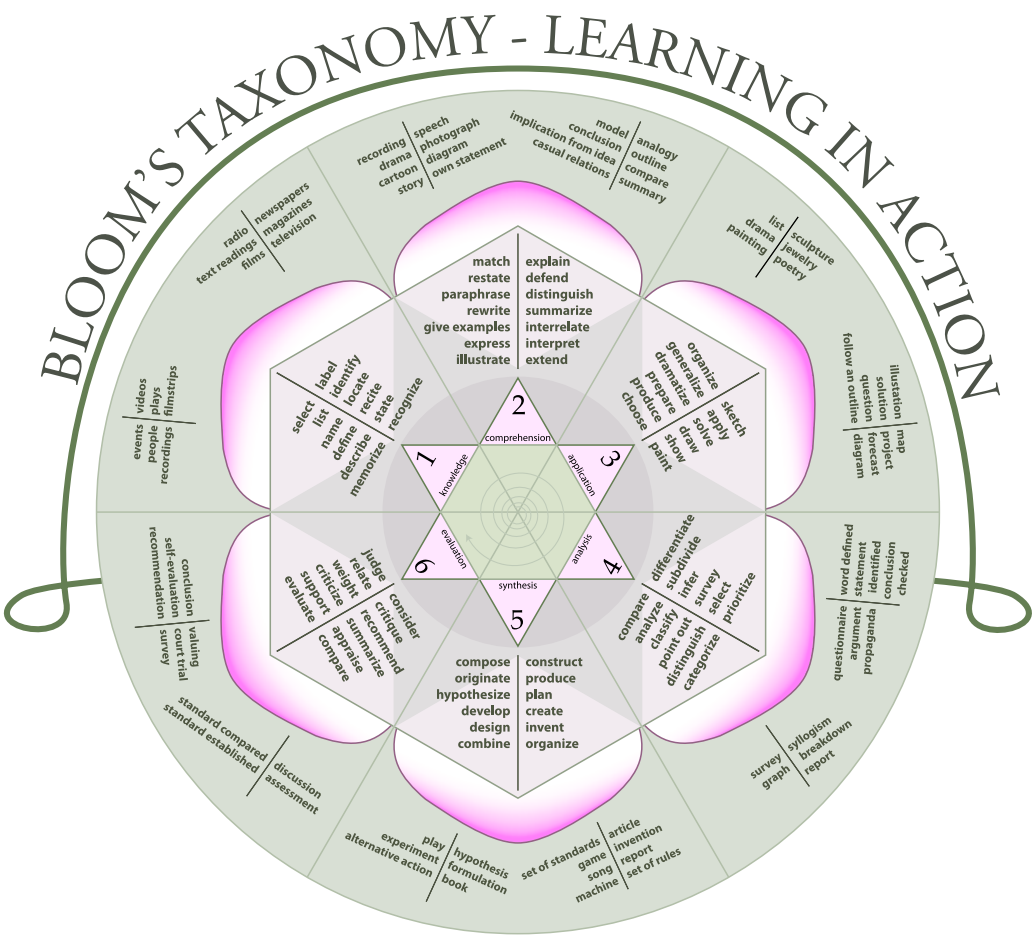

Illustration by K. Aainsqatsi licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License

Affidavit

I hereby declare that I am the sole author of this bachelor thesis and that I have not used any sources other than those listed in the bibliography and identified as references. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.

Place, date	Signature