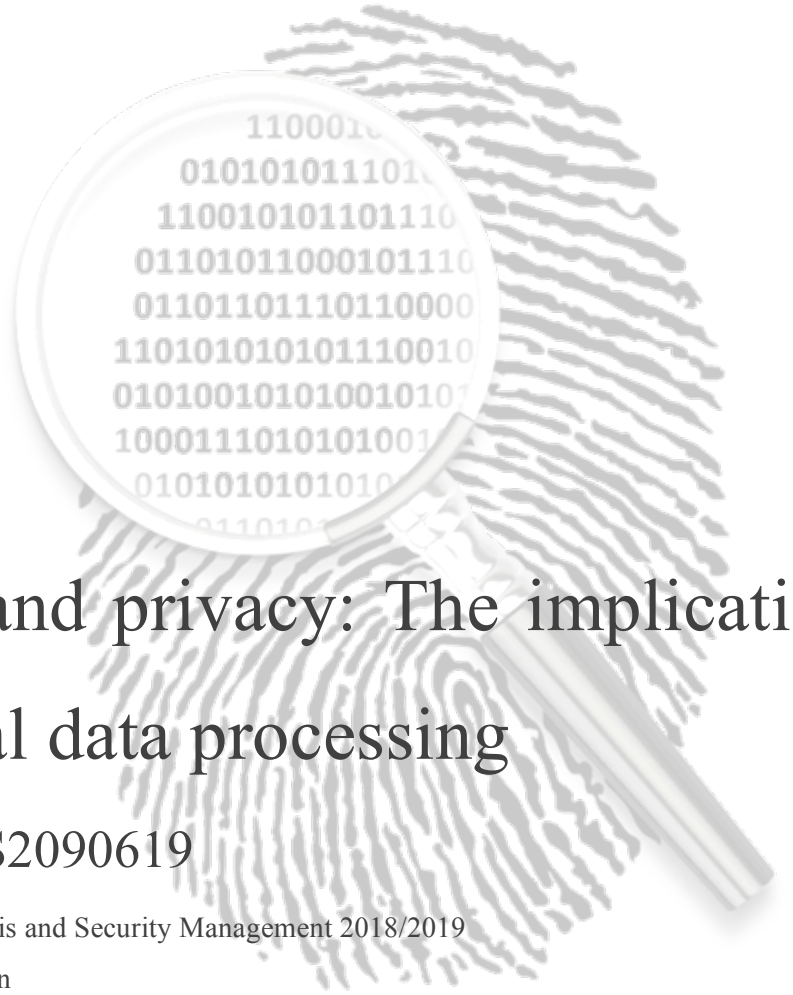
Universiteit Leiden
Governance and Global Affairs



Big data and privacy: The implications of personal data processing

R.M. Vlok – S2090619

Thesis for the master Crisis and Security Management 2018/2019

Supervisor – Dr. van Steen

Second reader – Dr. de Busser

Hand-in date: 09-06-2019

ABSTRACT

Big data is a phenomenon that has become increasingly relevant in the past decades as society generates increasing amounts of data. Large amounts of the generated data contain information about individuals. The processing of personal data is promising for organizations as valuable insights on individuals and groups in society can be found. Individuals are sharing increasing amounts personal data to get access to products and services. The purpose of this thesis was to explore whether the processing of personal data in big data environments can affect privacy. Privacy is defined as the right to live a life free from others, without unauthorized inferences, and to be in control over one's own data.

In order to explore this topic and answer the research question primary and secondary data have been collected through a literature study and by conducting interviews. The interviews have been conducted over a period of two months with experts in the fields big data and privacy from the Netherlands. In total nine interviews have been conducted which after being transcribed came to a total of 31.580 words.

The literature presents three challenges to privacy when it comes to processing personal data in big datasets. These three challenges are re-identification, targeting based on profiles and spurious correlations. The interviews presented that processing of personal data in big datasets has three upsides and four downsides for individuals. While these were examined considering the concept of privacy it is concluded that privacy can be affected by processing personal data in big datasets. Privacy can be affected as processing personal data in big datasets can result in unwanted interferences, persuasive pressures due to information limitation, loss of control over data and exclusion.

TABLE OF CONTENTS

INTRODUCTION

In August 2006, the American company AOL was in the news after publishing a large dataset containing 20 million web search queries by its users. At the time AOL was one of the largest internet providers in the United States. The dataset was released for research purposes and made public on the website of the company. Before publication of the dataset, the data that could be linked directly to individuals had been removed and replaced with numbers. For example, one user's pseudonym was No. 4417749, while another could be No. 3505202. This attempt to ensure anonymity proved to be weak as AOL failed to anticipate how unique online behavior is. With each click and each search query a user makes itself increasingly unique until eventually investigators were even able to identify users by their first and last name. When researchers approached a woman and got her permission, a woman named Thelma Arnold stepped forward and admitted that the search history of user No. 4417749 was her search history. She is just one of the 657.000 Americans in the dataset that could possibly all be identified. AOL removed the datasets days after publishing it but unfortunately the dataset had been downloaded by multiple users that were now re-uploading the set onto different internet portals (Barbaro & Zeller Jr, 2006).

A similar example comes from the on-demand streaming company Netflix. The company released an anonymized dataset that contained movie ratings of nearly 500.000 users. Participants of a contest, of which the aim was to produce an algorithm that could accurately estimate how much a user would enjoy a movie based on his or her preferences, were given access to the dataset. The grand prize for the contest that came up with the best algorithm was 1 million US dollar (Netflix, n.d.). Narayanan and Shmatikov, who did research with the dataset, showed that with little auxiliary information individuals could possibly be re-identified in the dataset. The Internet Movie Database (IMDb) is a website on which users voluntarily provide personal information and that allows for movies to be rated. Their user database proved to be perfect auxiliary information for the experiment (Narayanan & Shmatikov, 2008, pp. 12-13). Narayanan and Shmatikov found that with just eight movie ratings and dates, which allowed for two wrong ratings to be wrong and a 3-day error, 96% of the Netflix users in the dataset could be identified. Even based on only 2 ratings and dates the 500.000 in the dataset could be reduced to 8 people in the entire set with 89% certainty (Dwork, 2008, p. 8). This example may appear innocent as people may not care about who knows what movie they liked, but it illustrates how data that is perceived anonymous can be traced back to an individual.

In the digitalized world that we live in nowadays nearly every product or service collects data. Everyday humanity generates 500 million tweets, 70 million photos on

Instagram, and 4 billion videos on Facebook. It is estimated that 2.5 quintillion bytes of data are created every day (Calude & Longo, 2017, p. 2). Society enjoys the many benefits that come with product and service providers that know exactly what the consumer wants based on vast amounts of data. People enjoy the fact that digital advertisements display the products that they are interested in, or the luxury of their phones telling them how long the journey is to frequently visited locations. By putting such data in large data sets, combined with data from thousands of other individuals, the service providers can offer advice on what you might potentially like. This advice can for example be based on what other individuals, whom have similar habits, enjoy to use. Next to commercial use, large combined datasets are be used to recognize patterns, which may go unnoticed in regular data processing (Tene & Polenetsky, 2011). Profiles are created based on, for example, habits, preferences, geographical location and many more traits that can be used to categorize groups of people (Hasan, Habegger, Brunie, Bennani, & Damiani, 2013, p. 25). An example of how these profiles can be used comes from current US president Trump's election campaign. Trump is said to have used profiling to determine what parts of his campaign were especially important for individuals that fit within a certain profile in order to get them to vote for him (Gonzalez, 2017, p. 11)

One can wonder whether there are consequences to having personal data processed in this manner. The mass collection and processing of personal data bring challenges to protection of personal privacy. Can privacy be guaranteed if personal data is being collected and processed at such large scale? Can one remain on charge of his ability to take decisions? Does saved personal data pose any unobvious threats to privacy? Can individuals be influenced without being aware why they are being targeted? These are questions that come to mind when exploring this topic. The next section will present the main question for this research and sub questions, which will help answer the main question.

RESEARCH QUESTIONS

In order to shed light on the topic this research attempts to answer the research question that is as follows:

How can privacy be affected by processing personal data in big datasets?

With this central question it is crucial to explore the concepts of privacy, personal data and big data. The sub questions each help explore a part of the topic. These questions help test the hypothesis that is presented in the research design chapter. The sub questions are now introduced after which the rationale for each question is explained.

1. *What are the potential consequences of personal data processed in big datasets to privacy according to the existing literature?*

This first question explores the existing literature in the field of big data processing and its challenges to privacy. The literature that goes into this specific topic is scarce and mainly originates from the past two decades. Nonetheless, the exploration of previous studies will map out the academic landscape for this study. This question together with the concepts, is the part in which the existing literature will play an important role.

2. *Is personal data perceived differently between interviewees?*

Personal data is a concept that may be perceived differently between experts. What is considered as personal data may influence the way this data is processed, which in turn may influence its impact on privacy. This question will be answered based on the data gathered through interviews. During the interviews questions related to what is personal data create awareness that will be useful in the following sub question.

3. *What are the up- and downsides of personal data in big datasets for individuals?*

The third question is one that can be separated into two parts, the upsides and the downsides of personal data in big data. By considering the up- and downsides it became apparent what individuals are gaining by having their data processed as well as what consequences it can have. The answers to this question allowed for the examining of possible consequences of personal data processing to privacy.

4. *What are the possible effects of personal data processed in big datasets to privacy?*

Considering the answers from the previous sub questions this question will estimate the effects of big data processing to privacy. By knowing what is considered personal data, what the up- and downsides of the processing of this data in big datasets are, paired with interview questions on how this affects privacy, the possible effects will be estimated.

This thesis is organized in the following order. First, the concepts of this topic are conceptualized. This conceptualization will define workable definitions and give explanations of the practical application of the concepts. Secondly, the research design is presented. This includes the causal mechanism, methods used to collect data, the operationalization of the concepts and a discussion of the limitations and validity of the research. Thirdly, after the research design, the results are presented. The results from the literature and results from the conducted interviews are each presented in a separate chapter. After these two chapters, the results are interpreted in the discussion which allowed for confirmation or falsification of the hypothesis. Lastly, a conclusion is presented based on the conducted research and recommendations are given.

CONCEPTUAL FRAMEWORK

CONCEPTUALIZATION

The following subparagraphs provide an explanation on the main concepts of this thesis. These concepts are big data, privacy and personal data and are crucial for this research. The definition of each concept lays out the scope on this concept throughout the thesis, as well as what the concepts means in practice.

BIG DATA

In today's information driven society data plays a crucial role. Digitalization allows for major decisions to be taken based on large quantities of data. Before digitalization the collection of data was time consuming and the sharing could hardly be done in an efficient manner (Vetzo, Gerards, & Nehmelman, 2018, p. 14). Nowadays data is collected more easily through all devices and services used by society. Kitchin (2014, pp. 80-85) mentions that enormous increase in data collection has been made possible by the invention of the computer and the internet. Next to this the price of devices that save data has significantly decreased over the years. Devices connected to the internet allowed for part of our lives to be lived online. More data was collected in the year 2016 than in the entire history of humanity up until 2015 (Vetzo, Gerards, & Nehmelman, 2018, p. 14). These incredible amounts of data allow for decision making based solely on this data. Big data is a heterogeneous concept and can include any data imaginable. It encompasses anything from viewing habits on YouTube to details collected by medical appliances in hospitals. Big data offers possibilities to discover new relations between data points. In a report by the Dutch expert group on big data and privacy (2016, p. 11) they state that the power of big data is in the insights it creates through advanced models of behavior and techniques. The models recognize patterns and apply these in order to gain new insights in the preferences and behavior of individuals. Big data allows for characteristics of groups of people to be recognized in order to find relationships between possibly unrelated characteristics in order to gain insights in behavior (Expertgroep Big data en privacy, 2016, p. 11).

There is a variety of definitions for the term big data, most of these categorize big data on three characteristics often referred to as the 3V's. The 3V's are characteristics that are the most reliable indicators in order to categorize something as big data. These characteristics are Volume, Velocity and Variety (Torra & Navarro-Arribas, 2016). Volume refers to the huge amounts of data. There is no minimum size or amount of data set as a limit for a dataset to be considered big data. This makes it an ambiguous concept and open to interpretation. It is however characterized by collection of all data that belongs to a specific set. Meaning that it

should include all data that can possibly be collected on a specific topic. While traditional data analyses make use of limited amounts of data and attempt to generalize this over a population, big data is not limited because of the large quantities of data (Vetzo, Gerards, & Nehmelman, 2018, pp. 15-16) Velocity refers to the dynamic nature of big data. While traditional data is often collected at a specific point in time from a selected target audience, big data is collected real-time and can be acted upon in real time. An example of this are websites that show products based on the online paths taken by visitors (Vetzo, Gerards, & Nehmelman, 2018, p. 17). Variety refers to the variety of sources that is needed to realize big data. Big data comes from a large variety of sources such as social media, smartphone applications, government databases and other devices connected to the internet (White House, 2014, p. 5). Data collected in one set can be used in other areas as data is increasingly interconnected. An example of this is the fact that smart meters that measure electricity also save data on which brand of home appliances people use (Wetenschappelijke Raad voor het Regeringsbeleid, 2016). This data can thus be used for marketing purposes. In general the insights created by big data processing can be used to target individuals with specific recommendations and services, that can in turn influence their decision making (Expertgroep Big data en privacy, 2016, p. 11).

Besides information on individuals, big data can also include information that is not related to individuals in any way. Machine data such as data from sensors is not related to individuals (Supriyadi, 2017, pp. 30-31) but can be very valuable in big data analytics. Such data can for example point out weak sections of systems or show which parts are likely due for replacement based on big data analyses. This data is however irrelevant for this study as it does not pose a potential risk to privacy (Supriyadi, 2017, p. 31), and therefore excluded from the scope of this thesis.

The dynamic nature of big data and the fact that it is being fed by large quantities of available data make in a concept intriguing. A concept that needs to be studied more in order to fully understand its potential. Regardless of how thoroughly it has been studied, much is already happening with and based on big data. The following paragraphs will shed light on some events involving big data that have taken place.

BIG DATA IN PRACTICE

While processing big data is when its significance becomes apparent. While traditional analyses rely on the testing of hypotheses, big data does not require these. Big data allows for the recognition of patterns and connections over populations much larger than in traditional research (Wetenschappelijke Raad voor het Regeringsbeleid, 2016, p. 38). These patterns can be recognized as data from different sources is brought together. Similar data types can be connected, which results in comprehensive datasets.

A known example of the usage of interconnected datasets comes from former American president Obama's election campaign. Databases that held information on political preferences were combined with personal details such as music preference. As Obama invited supporters to join a dinner it became apparent that invitee's received a variety of invitations. Each invite included aspects of the night that might be of specific interest to this invitee (Crovitz, 2012).

Another example dates back to 2018, when Facebook was in the news in combination with the company Cambridge Analytica. It started in 2013, when researchers at the University of Cambridge were analyzing the data of people who completed a personality test on Facebook. This personality measures the metrics openness, conscientiousness, extraversion, agreeableness and neuroticism, which together form the acronym OCEAN. The population for this research contained 350,000 people from the United States. The data from the OCEAN-test was correlated with Facebook activity and showed a clear relationship between the two. The research demonstrated that the data from this OCEAN-profile could be deduced reasonably. Knowing that such analysis could be formed the, the Global Science Research cooperated with Cambridge Analytica and developed a similar personality quiz on Amazon's platform called "Mechanical Turk". This quiz required participants to give the access to their Facebook profile and to the users' friends' data. This gave Cambridge Analytica access to the data of tons of Facebook users. Cambridge Analytica soon realized that this data could be correlated with other types of data such as data on online purchases, browsers, voting and other social media platforms. The Facebook data combined with other public and private data allowed Cambridge Analytica to target individual consumers or voters with communication that could possibly influence their behavior. Such targeting was for example used in the election campaign of current United States president Donald Trump (Isaak & Hanna, 2018). Extensive profiles such as the OCEAN profiles enable organizations to feed individuals that fit a certain profile specific pieces of information.

While the examples above might come across as negative, the usage of big data can also have positive implications and possibly save lives. The discovery of the adverse effects of Vioxx painkillers can also be attributed to the use of big data. Kaiser Permanente made an analysis of the clinical and cost data and was able to identify that 27,000 cardiac deaths between 1999 and 2003 could be traced back to the usage of Vioxx painkillers (Tene & Polenetsky, 2011, p. 64). Permanente made this discovery by combining the datasets of clinical data and cost data and looking for patterns and correlations. Without this discovery there may have been more people suffering from the side effects of the painkillers. Another innovative example of big data usage is a service called "Google Flu Trends", which predicted outbreaks of the flu by using aggregate search queries. This was used to prevent epidemics by recognizing an outbreak very early on (Tene & Polenetsky, 2011, p. 64).

The mentioned examples are the tip of the iceberg of what can possibly be done with big data. The usage is however controversial as its usage for commercial purposes or usage in order to influence voting behavior can be considered unethical. Correlations in datasets can reveal information about individuals that they rather not openly share. Because of this the privacy of the concerned individuals is to be taken into account. Privacy is a concept that has changed over the past decades. In order to determine what it means for this study the following section lays out the definition of privacy.

PRIVACY

Privacy is a term that has become increasingly relevant in the information age. Society willingly shares more information about itself than it has done ever before (White House, 2014, p. 3). While individuals generate large quantities of information, it is necessary to think about where all this data could end up and what it could mean for one's privacy.

Multiple definitions of privacy exist as it is an ambiguous concept. The first mention of privacy as we still know it today comes from the study on the right to privacy by Warren and Brandeis. Already in 1890, Warren and Brandeis stated that new definitions of what is to be protected as privacy are needed from time to time (Warren & Brandeis, 1890). This remains true to this day as the concept of privacy reaches higher levels of importance due to technology evolving (Vedder, 2009). The definition as laid out by Warren and Brandeis is the right to be let alone. Originally this referred to being let alone from battery and assault but as time passed this definition developed and expanded (Warren & Brandeis, 1890, pp. 193-194). In the Netherlands it is commonly referred to as the right to have a life that is private from others and is closely connected to notions of human dignity and personal autonomy. Human dignity is described as the level of protection in regards to governments and third parties (Vetzo, Gerards, & Nehmelman, 2018, p. 53).

Anthony, Campos-Castillo and Horne (2017, p. 251), define privacy as the access of one actor to another actors information as well as the way information is used. Access can however vary in amount, type of access and content. A range of factors such as laws, social practices and technology affect the access to information. Laws referring to the level of access that is legal. Social practices meaning the levels of supervision and interaction patterns. Lastly technology, which refers to the systems that allow one actor access to the data of others. Next to technology, privacy norms may also affect access as they identify the characteristics of access that are socially accepted for the context (Anthony, Campos-Castillo, & Horne, 2017, p. 251). Anthony et al. give the example of it being acceptable to see nearly naked bodies on the beach but seeing the same bodies would be unacceptable through a neighbors' window. Violations of privacy norms is a combination of the level of access, the type of information that access is granted to, access through inappropriate channels and lastly

the inappropriate use of information. When privacy norms are followed, individuals feel that they have privacy. When norms are violated individuals feel invaded or isolated. These norms are determined by a variety of contextual factors, such as the relationship between the two actors, the purpose of the access and the way information is used (Anthony, Campos-Castillo, & Horne, 2017, p. 251). A definition of privacy, as given by Westin (1967, p. 7) is "the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information about them is communicated to others". Already in 1996 this definition was deemed more relevant than ever due to technological advancements in (Byford, 1996).

Burgoon et al. describe privacy as a multidimensional concept that can be explained through four dimensions, which are the physical dimension, interactional dimension, psychological dimension and the informational dimension. The first referring to intrusions to physical environment such as physical presence, sound, touch and odors. The second dimension, the interactional dimension, referring to the control of who, what when, and where we encounter others. The psychological privacy dimension describes the protection from intrusion to one's thoughts, feelings, attitudes and values. This includes freedom from persuasive pressures. The fourth dimension is the information dimension and refers to the ability to control who gathers and spreads information about one's self (Burgoon, et al., 1989, pp. 132-134). For this research, the definition of privacy is the right to live a life that is private from others, without unwanted interference from others, and the right to be in control of access to one's own data. The right to be in control of access to your own data refers to being able to determine the amount of access one willingly grants to others.

As described in the chapter on big data, this data is usually of such volume that it has to be processed through advanced computer systems. The focus of this research is on privacy in relation with big data. What is to be protected in order to live a life private from others in the context of big data, are personal data or personal identifiable information as this type of data can allow for interference with one's private life. The protection and care for personal data are important aspects to ensure privacy in today's information age. The next section will lay out a conceptual definition for personal data.

PERSONAL DATA

Personal data is a concept that can be broadly interpreted as it can be argued that personal data reaches much further than one may think. The definition in the legislation as presented by the European Union is as follows.

Personal data is defined in the Article 4 of the General Data Protection Regulation (GDPR), the EU's leading framework when it comes to data protection, as "any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as

a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;" (Council of the European Union, 2015). This definition is broad and sets a standard under which the individuals' data is to be protected when it comes to personal details. The GDPR also defines a list of special personal data, this data cannot be processed unless specified differently in a law. Special personal data are: information about race and ethnicity, religion, memberships of unions, genetic or biometric data, health, sexual orientation or sex life (Autoriteit persoonsgegevens, n.d.). In practice, personal data is what needs protection to ensure privacy. By having personal data unwillingly open for access by unauthorized others, privacy can be harmed as it may interfere with their right to live a life private from others.

Guidelines published by the Dutch Ministry of Justice and Safety state that for data to be personal data it should be about a person or be in regards to a person. The data should allow for identification of an individual. This can be through direct identifiers such as a name but also through specifics in one's appearance, such as length and hair color, or socio-economic characteristics, such as profession or income, and through online identifiers such as IP-addresses (Schermer, Hagenauw, & Falot, 2018, pp. 24-25).

A person is deemed identifiable if an individual could possibly be identified based on the available data. Even if no identification has taken place but it could reasonably be done, a person is considered identifiable. Identification usually happens though linking data to directly identifiable characteristics or by finding a combination in the available data that is unique enough to only refer one single individual. An example of the first situation is a phone number, which is considered indirectly identifiable, that can be connected to a name in a phonebook (Schermer, Hagenauw, & Falot, 2018, p. 25). An example of the second situation, described as spontaneous identification, is a 26-year-old public administration student living in the *Schouwburgstraat* in The Hague. This combination is so specific that it is very unlikely that more than one person fits this description. The following section will shed light on the usage of personal data in the context of big data analytics.

PERSONAL DATA IN THE BIG DATA CONTEXT

As illustrated in the two identification examples in the previous section, the legal definition of personal data is far-reaching. Data may not be used without permission if it can be linked to an individual. A solution that allows for the data to be used in big data analytics is the removal of personal identifiers (Supriyadi, 2017, pp. 31-32). Supriyadi (2017, p. 30) explains that given the legal definition of personal data, the definition of non-personal data is any data referring to a non-natural personal when such information does not convey any identification

of a natural person. This includes anonymous data namely information that does not relate to an identified or identifiable natural person. Anonymization thus refers to excluding personal identifiers from big datasets (Mayer-Schönberger, 2013, p. 142). Anonymization allows for the usage of non-personal data in big data analytics. Big data analytics are however tricky as the combination of various anonymized, and thus non-personal, datasets may result in identification of an individual (Supriyadi, 2017, p. 32).

The previous sections have conceptualized big data, privacy and personal data and explained its practical application. With the key concepts of this study explored they can be used in the causal mechanism. This will be explained in the following section, which presents the research design.

RESEARCH DESIGN

This chapter presents the research design of this thesis. The first section presents the hypothesis that is to be confirmed or falsified based on the outcomes of the research and a brief operationalization of the most important concepts. The second part of this chapter presents the methods that were used to collect data.

CAUSAL MECHANISM

For this exploratory research, the variables privacy and personal data saved in big datasets have been identified. This research attempts to explore how privacy can be affected by processing personal data in big datasets. Personal data processed in big datasets is the independent variable, while the dependent variable is privacy. Based on this the following hypothesis was defined and to be tested by this research:

Hypothesis 1: *"Privacy can be affected by processing personal data in big datasets"*

This hypothesis will be either confirmed or falsified based on the results of this research. The earlier presented sub questions helped gather the relevant results to test the hypothesis and answer the main question. The following section will explain what type of study is to be conducted and how the data is going to be collected.

METHODS

The data collected for this study is secondary as well as primary. The secondary data was collected through desk research, which laid the foundation for the primary data to be collected. The secondary desk research explored the current field of knowledge. Subsequently, a qualitative study in the form of semi-structured interviews has been chosen for this research as the topic is relatively new and the literature on this exact topic is limited. Interviewing was chosen as the method for primary data collection as the topic of this thesis is relatively new.

The desk research started with the literature written on de-anonymization techniques by Bruce Schneier. Schneier is an expert in the field of technological security. Through Scheier's literature the first articles on big data have been found. These articles highlighted privacy challenges of processing personal data in big data and were then further researched. Nearly all sources used to gather the secondary data are digital. These have been found through online services such as Wiley Online Library, Annual Reviews and Google Scholar. The majority of sources used, are journals articles and documents published by the Dutch government or research groups commissioned by the Dutch government. Literature has been selected based on its scope towards big data and privacy. While there are many things that can possibly affect privacy it is important for this research that the selected literature is on the

challenges and risks posed by big data. More specifically on the challenges and risks posed by processing personal data in big data and thus all literature was carefully examined and selected based on its significance specifically to big data and privacy.

Qualitative studies allow for a deeper understanding of social phenomena and beliefs. The purpose of qualitative studies in the form of research interviews is to explore views, beliefs and experiences of interviewees (Gill, Stewart, Treasure, & Chadwick, 2008, p. 292). The primary data was collected through semi-structured interviews that were conducted with individuals that have experience with big data and are knowledgeable in the field of privacy. The individuals were selected by means of non-probability sampling, as it is crucial that the interviewees have knowledge on the topic. Interviewees may have different experiences with big data and therefore semi-structured interviews are the chosen method. As described by Gill et al., a semi-structured interview defines key area's to be explored but leaves room for divergences in order to retrieve a more detailed response (Gill, Stewart, Treasure, & Chadwick, 2008, p. 291). Interviewees may not directly give a desired answer to a question and being able to probe deeper into a subject will allow for optimal results (Mathers, Fox, & Hunn, 1998, pp. 2-3).

Participants have been asked questions on certain themes. The first few questions were on the background of the interviewees, these are in place in order to categorize the responses and possibly these to different types of professions. After these questions the interviews moved forward into the subject matter. Questions exploring their perception of big data were to determine whether the perception of big data is the same for each interviewee. Differences in perception may have influenced responses later in the interview. The following theme focused on their knowledge of personal data in big data. Once again their knowledge on how personal data is processed in big data could show how involved the interviewees are in this topic and make them aware of how personal data is actually processed. This can determine whether they are actually aware of the protection techniques. The third theme is on the up- and downsides of personal data in big datasets. This answered the third sub question of the research and helped analyzing the possible consequences. In the last part of the interview the effects of personal data in big datasets to privacy were discussed together with the protection and awareness towards this topic. This part is crucial as it is directly linked to the main question of this thesis. The questions in this theme made visible whether the interviewees have the same or a different vision on the challenges to privacy as the literature. Do they see potential effects to privacy due to the personal data being saved in big datasets?

Ideally the interviews were conducted face-to-face. If this was not possible due to circumstances, interviews were conducted over the phone. The interviewees are from private corporations as well as governmental agencies. All interviewees are working with big data and were asked about the potential privacy concerns that big data processing brings. The

selected participants have thus been chosen based on their experience with, or knowledge of, big data and privacy. Examples or interviewees are researchers in the field of big data and privacy, data protection officers, experts that give public speeches on big data and privacy, big data consultants, and lawyers specialized in privacy protection.

The geographical scope of this research is on the Netherlands. Therefore, everyone that has been interviewed for this research is from the Netherlands and the interviews have been conducted in Dutch. The quotes presented in the result section have been translated to English. A total of nine interviews have been conducted. At the start of each interview the interviewees were given the same introduction. At the end of this introduction they have been asked if they agreed to the interview being recorded and whether their name could be used throughout the research. The data collected through the interviews was first transcribed after which it has been coded in order to be able to classify the responses. Appendix 1 contains the designed questions that were asked during the interviews. After all interviews had been conducted the responses were first categorized using comments. After all relevant information had been commented on, the responses were categorized into different themes using colors. A legend of which colors represents what theme can be found in appendix 2. Appendix 3 to 11 contain the transcripts of the interviews including the comments and colors used for coding. For publishing the transcripts have been excluded. The total word count of all interview transcripts combined is 31.580.

The next chapter presents a brief operationalization of the concepts of this study. After this the results of a deep dive into the existing literature on the risks of personal data in big data to privacy.

OPERATIONALIZATION

As described in the causal mechanism, the variables for this research are personal data processed in big datasets and privacy. The interviewees have been asked to define what they consider personal data and whether this data is a part of big data. The definition of personal data given by the interviewees is combined with the definition provided in the literature. By knowing what data is considered personal data, it can be estimated what consequences the processing of this data might have to privacy. This estimate is based on what the interviewees foresee as possible downsides for individuals, together with the possible consequences found in the literature.

Privacy is conceptualized as the right to live a life that is private from others, without unwanted interference from others, and the right to be in control of access to one's own data. The four privacy dimensions as explained by Burgoon et al. are taken into account to justify and categorize possible intrusions to privacy. In order to determine whether privacy can be

affected, the possible consequences are evaluated while privacy norms are considered. If these consequences can possibly affect the previously described right to privacy, it can be determined that personal data processed in big datasets can affect privacy. How privacy can be affected will be analyzed based on the given downsides and consequences of personal data processing in big data.

LIMITATIONS AND VALIDITY

The discussion of this study is based on the results gathered through literature and semi-structured interviews. Limitations of this method are that it is difficult to exactly repeat an interview, it is hard to generalize results and conducting interviews is time consuming (n.a., 2019). In order to ensure the highest degree of internal validity, and thus exclude other interfering factors, the abstract concepts are extensively conceptualized. In the discussion the focus is solely on the defined concepts and the potential consequences that have been found in the data collection. By focusing only on these consequences and their possible effect on privacy, other interfering factors are disregarded. Privacy is a broad concept and in order to able to determine whether it can be affected its definition for this research has been tightly defined. A threat to internal validity using the chosen method of interviewing is the fact that interviewees may be influenced in their answers by previous experiences. Especially due to the experts coming from different disciplines, their own interpretations of the concepts may vary. It was necessary to conduct interviews with experts from different backgrounds as the amount of experts in the Netherlands on both big data and privacy is low. In order to minimize interference based on different interpretations of concepts the interviewees are asked to define some of the concepts based on their background. The total amount of interviews conducted did not lead to a point of saturation in information. Saturation could not be reached due to limitations in the amount of available experts on the topic and due to time limitations. It became obvious that saturation was not reached as up until the last interview new information was brought up. The fact that total saturation has not been reached leaves open possibilities for other possible consequences that may not have been found in this research. Even though conducting and processing interviews is time consuming, it has been attempted to conduct as many interviews as possible in the time given.

In order to ensure external validity, all interviewees have been sent the same e-mail to request the interview and have been given the same introduction upon starting the interview. During each interview the same themes have been discussed, questions did vary as the unstructured interview method allowed for probing into certain topics. As the interviewees are from different disciplines, some proved more knowledge able on certain topics. As this research focusses on the Netherlands and is based on the expertise of Dutch interviewees the

results may vary if the same research is conducted in another country. Definitions of the concepts could differ based on the legal frameworks of the country at hand.

With the causal mechanism, methods, operationalization and validity presented the research design is concluded. The following two chapters present the results gathered through the described methods. First the results from the literature are presented and concluded. Secondly, the interview results are presented.

LITERATURE

Research conducted in 2009 examined the size of social media users' online social footprint. Online social footprint is a term that describes the amount of profile's that can be linked directly to an individual and the amount of fields this individual fills in about themselves. This footprint is used to characterize a user's social networking activities. In 2009, Myspace and Facebook had around 250 million accounts, the average active member already had 5.7 accounts linked to their identity (Irani, Webb, Li, & & Pu, 2009, pp. 1-2). In the fourth quarter of 2018, Facebook alone had 2.32 billion monthly active users (Statista, n.d.). The online social footprint research has not been conducted recently but one can imagine how large our online social footprint is 10 years later. This is only information that we share willingly. Besides the information that we provide willingly there are vast quantities of information collected that can be considered personal data such as activities and living patterns.

By consciously and unconsciously sharing large parts of our lives it should be considered whether this could have implications to private life. The following sections attempt to find answers to the first sub question of this thesis which is "What are the potential consequences of personal data processed in big datasets to privacy according to the existing literature?".

CONSEQUENCES ACCORDING TO LITERATURE

As the chapter that defined big data already demonstrated, its potential is immense. Traditional research design is less relevant in big data analytics as it draws on much larger datasets. The data in many of these sets includes personal data. Datoo (2017) describes that in many cases personal data cannot be separated from the non-personal data. The mass collection of personal data and storing it in big datasets bring challenges in regards to data and privacy protection in big data analytics. As was ruled in the lawsuit commonly referred to as the "Digital Rights Ireland Case", the retention of personal data directly affects the right to privacy when the data allows for conclusions to be drawn concerning the private life of the persons whose data is processed, such as on habits, places of residence, daily movements, social relationships and social environments (Bredenoord, van Delden, Mostert, & van der Sloot, 2017, p. 6). The following sections will bring forward the risks that are presented in the existing literature.

RE-IDENTIFICATION

As upon processing data can directly or indirectly identify an individual, the metrics that allow for identification are often removed. Parties that process data use such techniques

to protect the privacy of the individuals involved. Tene and Polenetsky (2011, p. 65) state that traditionally organizations use de-identification techniques to hide real identities of data subjects. Anonymization techniques worked well until about two decades ago (Ohm, 2009, p. 1716). Anonymization techniques aim at maximizing privacy by removing personal identifiers without harming data utility (Ohm, 2009, p. 1754). A thoroughly anonymized dataset without any identifiers and with maximum data utility can be considered the holy grail of big data analytics as this could be used for any type of analysis and still produce valuable insights.

Research by computer scientists has however shown that anonymized data can be re-identified and point out individuals (Tene & Polenetsky, 2011, p. 65). A study on a dataset released by the Group Insurance Commission in Massachusetts showed that matching information on 135.000 patients with simple demographic information gathered by a voter registration list allowed for successful re-identification. Some of the fields, such as date of birth, zip code, and gender, were present in both datasets and allowed for re-identification (Cavoukian & El Emam, 2011, p. 2). An American landmark study showed that 87% of the American population could be uniquely identified based on their ZIP code, birthday and year and sex (Ohm, 2009, p. 1705). The expert group on big data and privacy also describes that aggregated data, that does not refer to an individual, connected to other data can lead to re-identification (Expertgroep Big data en privacy, 2016, p. 13). By linking the entries from various datasets, the uniqueness of created profiles increases and can ultimately lead to identification.

Jensen (2013, pp. 236-237) distinguishes three types of re-identification attacks, these are correlation attacks, arbitrary identification attacks and targeted identification attacks. Correlation attacks refer to two datasets put together in order to see matches in in specific metrics in order to single out an individual. Arbitrary identification attacks refer to trying to match one entry in a dataset to an individual's identity to confirm their identity. Targeted identification attacks are attacks to find details of a specific human being and not from any random individual in the dataset.

Tene and Polonetsky (2013, pp. 251-252) argue that researchers draw very different conclusions from strings of online search queries. The example given that is very different conclusions can be drawn from the search query "Paris", "Hilton" and "Louvre" compared to "Paris", "Hilton" and "Nicky". Adding more and more queries to one's digital search profile it can become increasingly revealing. Tene and Polonetsky continue to argue that once a string of clicks is linked to an identified individual this becomes very difficult to disentangle. As soon as one piece of data can be linked to a person's real identity, any association between this data and a virtual identity breaks the anonymity of the virtual identity (Narayanan & Shmatikov, 2008). Ohm compares these combined search queries with a human fingerprint

left at a crime scene. A fingerprint can be linked to a single individual and link that person to more available information. A so-called data fingerprint allows identification based on combinations of values of data that are shared with nobody else (Ohm, 2009, p. 1723).

The previously given examples of re-identification of individuals such as re-identification based on movie rating may come across as harmless. These harmless cases do however make future re-identification easier. Technological advances also increase the utility of data and therefore databases can never be perfectly anonymous. What is anonymous now may with be re-identifiable in the future (Ohm, 2009, pp. 1705-1706). Already in 2009 Ohm described a hypothetical "database of ruin", which refers to the fact that every person in the developed world can be linked to a fact in a database that can be used to blackmail, discriminate or harass this individual (Ohm, 2009, p. 1748).

Re-identification of supposedly anonymous data is a direct risk to privacy as it may reveal information about individuals that they would not have wanted to share with anyone. The re-identified data can cause harm or difficulties to the individuals' private life. Secondly, the data can end up at non-authorized data processors, which in turn can use the data for activities that may interfere with one's privacy. This can for example be in the form of targeting based on profiles. The next subparagraph explains how this possibly affects privacy.

PROFILING

Analyses of data about consumers allow them to be targeted based on how much they fit within a certain profile. Data used to determine such a profile is refreshed often to keep it accurate. As data is processed over a longer period of time the accuracy of the predictive value of the models and algorithms is likely to increase (Expertgroep Big data en privacy, 2016, p. 15). The data collected is often used for customization and personalization of digital environments and content. Profiling is defined by the Dutch privacy Watchdog *Autoriteit Persoonsgegevens* (2018, p. 7) as the automated processing of personal aspects with the goal of predicting professional accomplishments, economic situation, health, personal preferences, interests, reliability, behavior, location or to predict physical movement. Profiling can be beneficial for users as it allows for efficient usage of services and giving accurate recommendations. For example, users on Netflix and Bol.com are shown recommended movies and products based on previous interactions. The same goes for Google's search engine and autocomplete functions (Polonetsky & Tene, 2013, p. 4), another example of a recommender system is Facebook showing potential friends. Such recommendations are made based on profiles, which consist of various attributes that may describe a user. Attributes may include geographical location, professional background, interests, preferences, opinions etc. (Hasan, Habegger, Brunie, Bennani, & Damiani, 2013, pp. 25-26).

To maintain accurate, the data needs to be refreshed often, this requires the people whose data is being collected to be increasingly transparent in their daily life. This in contradiction to the data collection itself which often is not very transparent. Categorizing individuals based on profiles created through big data analytics can exclude individuals or target them with products and services that they are not comfortable with. It could introduce them to products and services that they are entirely uncomfortable with, without knowing why this is shown to them, or exclude them without them knowing why they have been excluded. This can be experienced as an invasion of privacy. Based on profiles created through big data analysis individuals may not be eligible to buy a certain product or service, or under a different set of conditions. Many may not understand why they are excluded from a certain group or profile and unwillingly pushed into a different direction. Profiling can be seen as a direct consequence to privacy as it can cause one to partially lose control over the freedom to take decisions. As the processing of personal information in the service industry keeps growing it is likely that in the near future individuals will increasingly wonder how a service provider got these details about them (Expertgroep Big data en privacy, 2016, p. 15).

SPURIOUS CORRELATIONS

The Dutch expert group big data and privacy points out that the result of the analysis is not correct, as a statistical relationship does not necessarily indicate a causal relationship. The use of incorrect or outdated information has potential to cause problems when used to create profiles (GDPR Report, 2017). The example is given that the total revenue generated by arcades has a 98,5% correlation with computer science doctorates awarded in the US. This correlation is high enough that one may assume a relation between the two while the two variables are completely unrelated. Another example given is the correlation of 99.7% between the US spending on science, space and technology and the amounts of suicides by hanging, strangulation and suffocation. Insights generated through big data analytics may cause one to think these two are related while once again they are not in any way (Expertgroep Big data en privacy, 2016, pp. 15-16). Jensen refers to the fact that drawing conclusions based upon correlations made in datasets can pose challenges to privacy. These conclusions may be based on data linked to wrong individuals and therefore be entirely untrue (Jensen, 2013, pp. 236-237). This can be due to manipulations in the dataset because of unwillingness of data sharing or faulty interpretations of the data at hand (Jensen, 2013, p. 238). It is described that groups in society can be sorted based on correlations found in big data, which is referred to as social sorting. As these correlations are often not causal, it is not without risk to make conclusions based on these correlations. The demarcation of the created groups can be biased by basing the conclusion on a spurious correlation. If the data is seen and used as a perfect reflection of the group it will generate conclusions that do not fit the

group at hand. If the bias is not detected it can reproduce itself and become increasingly discriminatory (Wetenschappelijke Raad voor het Regeringsbeleid, 2016, p. 89).

It is more likely for spurious correlations to occur in big datasets compared to the traditional statistical research method. Big data powers data-drive analyses and is not about testing hypothesis but about finding correlations and patterns. Analyses based on big data can find correlations between any type of available data and causality between variables should be doubted (Wetenschappelijke Raad voor het Regeringsbeleid, 2016, p. 38). Considering the previous section on targeting individuals based on profiles spurious correlations can influence individuals as they will be placed in a category in which they not belong. Again, this can result in the individual being excluded, wrongfully targeted or influenced with content that is entirely irrelevant for this individual.

PUBLIC OPINION TOWARDS PRIVACY PROTECTION

In 2015 the European Commission conducted 1008 interviews with Dutch citizens on data protection as a part of the Eurobarometer. The report on this research shows that 65% of Dutch citizens are worried that information collected for one goal may be used for other purposes such as direct marketing, personal advertisements and profiling. 9% of the total amount of interviewees feels that they have total control over the information they share online, while 59% feels that they have some control and 30% feels they have no control (European Commission, 2015).

Sharing personal information is an increasing part of modern life. 58% of the participants stated that there is no alternative to sharing personal information if one wants to use products or service. 48% of the total does not mind sharing personal information, while in another question 60% does not like to share personal information in return for free online services. When it comes to trust in different types of organizations the interviewees trust healthcare organizations the most (81%), while online organizations such as search engines and social networks are trusted the least with 18% (European Commission, 2015).

In 2019 the Dutch privacy watchdog *Autoriteit Persoonsgegevens* published results of their research into the public opinion towards privacy. This research shows that 94% of 1002 participants worry about their privacy, while 1 in three worries a lot. Participants worry the most about online retailers, tech companies and banks and insurance companies. The top three fears are abuse of their data, unauthorized access and data in the wrong hands. 88% of participants has never used their right to privacy because they do not know how, think it is too much of a hassle or do not find it important (Autoriteit Persoonsgegevens, 2019).

CONCLUSION LITERATURE

The literature shows that people are generating increasing amounts of data about themselves. In the big data context it is challenging to protect this data and therefore privacy. As much data is being stored it is hard to separate the personal data from non-personal data and the retention of this data is a threat to privacy. The literature presents three challenges to privacy which are re-identification, profiling and spurious correlations. Firstly, techniques to de-identify, or anonymize, the data are used to protect the privacy of individuals. Such techniques are discredited in the literature, as individuals can be re-identified based on seemingly non-personal identifiers. Data that is appropriately de-identified poses possible risks to privacy in the future, as more advanced systems are possibly able to de-anonymize it. It is described that for each individual some information exists in a database that could be used to blackmail this person. The second challenge refers to profiles based on user behavior. Individuals can be targeted based on their own behavior and the behavior of individuals similar to them. This can expose individuals to targeting that they are uncomfortable with and one may not know why he is being targeted. The third challenge refers to spurious correlations that are found through big data analytics. As this type of analysis is data driven and combines all sorts of data in order to find correlations it has potential to find correlations which are not causal in any way. Individuals can be targeted based on conclusions drawn from spurious correlations which can be discriminatory and experienced as a violation of privacy.

Research by the Eurobarometer and the *Autoriteit Persoonsgegevens* shows that individuals are worried about their privacy and feel that they are not in control over their own data. Sharing data is seen as a part of modern life and is rather not shared in return for free products and services. Remarkably, trust in online search engines and social networks is the lowest, while these are possibly the largest collectors. Though trust is low and people worry, the large majority has never acted upon their right of to privacy. The next chapter will present the primary results gathered through conducting interviews.

QUALITATIVE ANALYSIS

The results presented in the following sections have been collected through interviews with experts on subject matter. After the results are presented they are linked to the explored literature where possible in the discussion chapter.

DEFINITIONS

In order to determine whether the interviewees have the same vision of what is considered big data, privacy and personal data they have been asked to define them. Generally, the definitions were similar to the ones defined in the chapter on conceptualization. There were however some discrepancies and additions in regards to the definitions. The following sections will introduce these.

BIG DATA

Big data is a term that proved challenging to define for the interviewees. As presented in the conceptualization, volume, velocity and variety are commonly seen as the characteristics of big data. Three of the nine interviewees mentioned these characteristics when describing big data. Two of these three added additional characteristics which are veracity, validity, visualization and value.

> *There are the well-known three V's. Volume Variety and Velocity (...) I once thought of the 7 V's, these included Veracity, Visualization, Value and, [hesitating] there is a seventh V. – Founder Datafloq – Appendix 11*

> *In the literature I have seen them [characteristics of big data], the 6 V's or 5 V's. Volume, Veracity, Validity, Velocity, Variety and another one. – Speaker on new technologies – Appendix 10*

All interviewees agreed that big datasets can contain any type of data. Besides this it is mentioned that big datasets are of such volume that traditional tools are not capable of analyzing the data. Next to characteristics that describe big data, it is referred to as a way of analyzing data in order to find correlations. Examples of this are given in the following quotes.

> *I will first explain small data in order to explain what big data is. Small data is when you want to research a phenomenon but the phenomenon is too complex, and too big to collect all available data. What you do then is work with samples, explorations and*

those kind of methods. The whole idea behind big data is that instead of working with a sample, you work with the data that is available. The enormous amount of data that is available is often messy or less organized. Not designed like in a sample experiment. (...) Big data takes the messiness for granted and use all available data to make an analysis. (...) Big data allows for the discovery of correlations that would otherwise not have been found. (...) You could attempt to find all possible relationships in a dataset and try to see if these correlations have meaning through qualitative research. (...) I would say big data is more a way of working than it is a type of data. – Founder Utrecht Data School – Appendix 8

Big data is when you have a large amount of different types of data, about purchasing behavior, income, health and in this large amount of data we are going to look for correlations. (...) You are trying to slice through the datasets in order to find relationships between data points in to order to come to conclusions on this and to undertake action based on these conclusions. – Manager Security and Data Protection Officer – Appendix 4

Big data is about datasets that are too big to analyze them with traditional tools. – Founder Datafloq – Appendix 11

The goal of big data analytics is to generate profiles that can be used to take decisions and undertake actions. These profiles can be on humans but also on situations and machinery. The founder of Datafloq explains that the use of big data can be split up into three parts. Firstly, the customer as you can build extensive customer profiles. The second part refers to the product that can be made perfect based on data. The third part is predictive maintenance on products or organizations, as data can show when specific parts need maintenance.

[Big data can be used for] different areas, I separate them in three parts. On one side the customer, you can get a 360 degree customer profile and discover new markets. (...) On the other hand the product, you can offer a personalized product. Which you can offer on the right place, through the right channel, for the right price. You can also do predictive maintenance on your product or organization. – Founder Datafloq – Appendix 11

The interviewed coordinator of a business intelligence center stated that big data gives better insights. It makes exceptions visible and helps calculate a weighted arithmetic

mean. The example is given of an organization that calculates creditworthiness of organizations.

> *It [big data] gives better insights. It makes exceptions visible to the users. (...) You can come to a weighted average. I worked at [company], which does credit ratings for about three million companies in the Netherlands. We then calculated the average credit risk of all cobblers in the Netherlands. So with the big data we came to an average which could be used as a benchmark. – Coordinator business intelligence center – Appendix 7*

Big data can be used to calculate an average that can then be used as a benchmark. Professor Law and Digital Society explains that big data is often about making decisions on individuals. This can be whether a consumer is creditworthy or to determine if someone it fit for a job or not. Such decisions can be made based on personal data but also on other non-personal variables that can be used to determine a profile, such as hardware used. If, for example, analysis shows that individuals using a Russian keyboard in combination with other attributes are more likely to be fraudulent, they could possibly be refused service.

> *It [the purpose of big data] is often about making verdicts on individuals. This can be on whether a customer is creditworthy or not, or whether he is fit for a job or not. – Professor Law and Digital Society – Appendix 6*

> *People are looking for correlations and patterns that have meaning based on all sorts of available information. It can be, while not always based on the processing of personal data, that an online shop, a web shop so to say, based on an algorithm of a third party decides not to service individuals that use a specific device, mobile phone or tablet or laptop that has a Ukrainian IP-address or uses a Russian keyboard or other variables. It could have been determined that this together has a higher chance of fraud. So in that case, if someone visits the web shop with those attributes, they may not accept credit card and payment has to be made in a different way, upfront. – Professor Law and Digital Society – Appendix 6*

Chief Privacy Officer (Appendix 3), gives the example that after the bombing during the Boston Marathon in 2013, a profile based on search queries could be determined. He continues to describe that after the bombing a man living in New York did a Google search for a backpack, the woman for a pressure cooker and the child for a third variable. All of these aspects were present in the bombing and based on a determined profile the FBI decided

to raid the house, only to find an innocent family. The founder of Utrecht Data School explains that big data can be used to make predictions and to categorize people, groups or things. These categorizations are made in order to approach people as efficient as possible with only relevant content (Appendix 8).

> *Making predictions is one [goal of big data]. The other is categorizing. You can say, you can categorize people, or groups or things in its broadest sense. – Founder Utrecht Data School – Appendix 8*

PRIVACY

Privacy is generally perceived similar to the definition in the conceptualization. The concept evolved from being free from intrusion, to being in control over your own data. The definition of privacy is ambiguous but is mainly about the inviolability of one's private life. It has been noted that privacy and data protection are two different things that are in close connection to each other.

> *Privacy goes back to 1948, the declaration of the Universal declaration of human rights, Article 12. It is about the inviolability of private life. As citizen you should be free from intrusion. (...) This is the basis of which we later started to call privacy. – Manager Security and Data Protection Officer – Appendix 4*

> *Privacy and data protection are two different things that are in line with each other. Privacy is that your neighbor does not look over the fences into your garden while you are sunbathing. By doing so he violates your right to privacy without violating the general data protection regulation because that regulation is on the processing of personal data. (...) For me privacy in essence is the right to be in control of your own data and being able to make well-considered choices without that choice being made by someone else while you are unaware of this. (...) To be in control of your own data, for me that is a very important core value of privacy. – Chief Privacy Officer – Appendix 3*

The owner of Datafloq explained that privacy is about the control that a person has over its own life and its own data. Individuals should decide who gets access to what information and when can this occur. This does not mean that personal data is not publicly accessible, as long as the individual chose this option that is fine.

I think privacy is the control you have as a person over your own life, your own data. Your information. You can decide who can see what information at what moment. Privacy does not mean that you information and knowledge is not publicly accessible. If you chose for it to be that is fine. Founder Datafloq – Appendix 11

The Chief Privacy Officer states that big data and privacy are at odds with each other. He explains that privacy is about surprise minimization while big data is about surprise maximization.

Big data and privacy are at odds with each other. Look, big data is about… Privacy is about surprise minimization and big data is about surprise maximization. They are inherently at odds with each other. – Chief Privacy Officer – Appendix 3

PERSONAL DATA

Personal data is a term that proved to be the most discrepant. The general consensus is on the fact that data can be classified as 'personal' when it can be traced back to an individual. It depends on the goal of the analysis at hand whether personal data is crucial in big data. Analyses that examine machine efficiency do not require personal data (Chief Privacy Officer, Appendix 3). Multiple interviewees mention that combinations of data generated by using computers and the internet can be considered personal data. It is mentioned that the definition of what is personal data is being stretched continuously. Examples of responses given during the interviews are given below.

We are constantly broadening the definition of what is personal. – Founder Utrecht Data School – Appendix 8

When I conduct a survey among a thousand employees (...) an opinion of a person or answers about his behavior to questions we asked. I consider that very personal. The law would not classify this as personal data. This data is not his name or address but a question would be whether he locks his computer before leaving his desk. (...) However, if it cannot be traced back to an individual it is no longer personal data. – Developer security programs based on personal behavior – Appendix 5

Personal data is all data that can be connected to one person. That can go very far, of course your address is personal data. But data on locations can be extrapolated and be personal data too. Research has shown that if you have enough data points of

a person, you can point out who that person is. This makes anonymous location data personal data as well – Founder Datafloq – Appendix 11

[Personal data] is everything that can be traced back to a person. This can be purchasing behavior. This can be a number plate on a car. It can be travelling behavior. Everything. It can even be the IP-address of my iPad because I always access the network through the same IP-addresses at night. – Coordinator Business Intelligence Center – Appendix 7

The definition of personal is of course very difficult. If I click on a website and somewhere it is registered which site that is, and I do that 5 more times, I have probably created a pretty unique profile of myself. Purely because I visited 5 websites in a combination that probably no one else has ever done before. […] And probably, no definitely, if I do that more often you can retrace what personal details are connected to those sessions. – Found Utrecht Data School – Appendix 8

Because you use an IP-address and cookies are placed on your computer. (…) Because they are "fingerprinting" they know that you use a Mac computer, this makes a combination of data. (…) This results in you not being directly identifiable but it does make you indirectly identifiable and when one can be indirectly identified it is considered personal data which makes the GDPR applicable. – Chief Privacy Officer – Appendix 3

Founder of Utrecht Data School (Appendix 8) argues that by linking various anonymized datasets, individuals can be identified relatively quick. Manager Security and Data Protection Officer (Appendix 4), explains that based on three characteristics 87% of all American citizens can be uniquely identified. These characteristics are zip code, gender and date of birth. The interviewed Professor at Law and Digital Society (Appendix 6) mentions that anonymization can be seen as a magic trick and that should not be trusted too much. The removal of first name, last name or date of birth is not enough as other combinations can lead to one person. The Chief Privacy Officer (Appendix 3) states that by combining anonymized or aggregated datasets individuals can be identified.

[Anonymization of personal data] that is happening yes, the point is, there have been quite some research projects into how good we are in anonymization. (…) If you put in a little effort and link datasets you can re-identify names and numbers based on anonymized data– Founder Utrecht Data School – Appendix 8

I have doubts about anonymizing data. It is sometimes seen as a magic charm. A characteristic for magic charms is that they are mythical or magic. Fairytales. Maybe we should not put too much trust in it and find out under what conditions it [big data processing] can be done. – Professor Law and Digital Society – Appendix 6

I do not believe in anonymization, especially not in big data environments. Anonymization is hard, so you really have to know what you are doing. When you have a set of data in which the personal elements have been removed, then the original set must be destroyed so that the connection can no longer be made. – Chief Privacy Officer – Appendix 3

To briefly conclude the interview results on the concepts of this thesis it can be said that even for experts the concepts big data and personal data are ambiguous. While some experts describe big data by its characteristics such as the three or more V's, others see it is a way of working with data. When considering what is personal data the experts acknowledge the difficulty of defining this concept. Data on behavior and clickstreams online may not contain any personal identifiers but are considered personal data by interviewees. In the next theme the experts have been asked about the up- and downsides for the processing of personal data in big datasets. The next section presents the results of these questions.

UP- AND DOWNSIDES

Big data is a phenomenon that, as previously mentioned, offers many possibilities for organizations. Datasets that offer these possibilities often contain data about individuals. This section presents the up- and downsides for individuals according to the interviewed experts.

On the upsides for individual the interviewees were generally likeminded. According to the majority of the interviewees, the biggest upside for individuals is content that is personalized based on their behavior, and the behavior of similar individuals. For example, the interviewed developer of security programs (Appendix 5) mentioned that through big data the ease of use increases. Next to the personalized content and the ease of use it has been mentioned that individuals can be offered a product that suits them perfectly and possibly a discount based on their profile. In the four quotes below show examples of how different interviewees described the upsides for individuals.

[Are there upsides for the people whose data is being processed?] Yes, I think that that is part of the problem There are many upsides to the processing of data. The

ease of use increases. – Developer of security programs based on personal behavior– Appendix 5

Of course there are upsides, better advice in content that you are shown. When an insurance company can better estimate what insurance you need, a part of the Netherlands will enjoy the benefits of this. These are the people that are healthy and doing well in life, at least doing well in life according to the norms of the insurance company. A good amount of people will profit from this. – Founder Utrecht Data School – Appendix 8

It could have upsides, maybe it has been determined that someone fits a profile can be give a certain discount because they are very reliable. – Professor Law and Digital Society – Appendix 6

The upsides are that as a person you get offered a better product, through the right channel, at the right moment, at the right price. (...) As a customer you get a better product and will not be bothered by irrelevant offers. – Founder Datafloq – Appendix 11

The downsides for individuals vary more than the upsides. One of the mentioned downsides is that organizations may use the processed personal data for other purposes than originally consented on by the individual. Individuals will hardly be able to do anything about this (Found Datafloq, Appendix 11).

The downsides come up when a business sells the data without permission. The downsides come up when businesses do not work according to the agreements made with the customer. Founder Datafloq – Appendix 11

The interviewed speaker on new technologies and Biohacker (Appendix 10) describes a situation in which based on online search habits Google can determine that one is likely to develop a depression. Google is not obliged to make this known to the individual but could decide to share this information to insurance companies. Based on this information one could be denied insurance while it would not be clear why one is being denied. The Professor at Law and Digital Society (Appendix 6) expresses worries towards people having the feeling that they are not in control of their own life and that decisions are made which they do not understand and are out of their control. He continues by saying that exclusion based on data is very close to discrimination. The developer of security programs (Appendix 5) states that

organizations can be too dependent on outcomes of big data analyses and that this may result in excluding people while it is not clear for the individuals why they are excluded. Organizations that fully trust the data can develop discriminatory models to work with. Often big data and the methods of calculation behind it are seen as the ultimate method but this can be questioned. The founder of Utrecht Data School (Appendix 8) explains that decisions on whether one can rent a car or can get a mortgage, can be made based on risk profiles while these profiles are no longer based solely on traditional information such as income.

> *[Name] is a neuroscientist and claims that based on smartphone behavior he can determine whether you are at risk to get a depression or another mental condition. Imagine that Google develops such an algorithm and based on that algorithm can tell me that I have a high chance to develop a mental condition. Google is not obliged to tell me. They can even sell the data to my insurance company or another insurer. Based on this I can be denied service and for me it is not clear based on what information this has happened. – Speaker on new technologies and biohacker – Appendix 10*

> *You just do not know that your details are being processed in analyses so you can hardly protest against it. You are not aware how certain conclusions have been made so you can be treated in a way without being able to do anything about it. – Chief Privacy Officer – Appendix 3*

Besides exclusion based on determined profiles another downside for individuals is the fact that they can be influenced and pushed into a certain direction. Filter bubbles due to algorithms can be problematic as individuals only get to see information that is close to their own vision and opinion. The following two quotes relate to this phenomenon.

> *One of the downsides can be that you get certain offers of products and content that push you in a certain direction. Or that do not expand your horizons. – Founder Utrecht Data School – Appendix 8*

> *If you have an individual that you can classify as someone with a gambling addiction based on attributes, and show this person advertisements for online casinos you can influence his life deeply. – Professor Law and Digital Society – Appendix 6*

The Founder of Datafloq (Appendix 11), explains that as long as organizations are only using data for the purpose that was agreed upon that the downsides for the individual are

kept to a minimum. Downsides exist when an organization is hacked and the data becomes publicly available or when the data is sold without consent. Manager Security and Data Protection officer (Appendix 4) states that the risks of current applications are often exaggerated.

> *It depends on what is done with the data. If you are a trustworthy organization and only use data for the purposes that the customer agreed upon. (…) Then there are not many downsides. (…) The downsides occur when a company is hacked and the data becomes publicly available, unprotected. – Founder Datafloq – Appendix 11*

It is mentioned in the interviews that big data analyses aim at finding correlations between different types of data. Chief Privacy Officer (Appendix 3), describes the risk that correlations are not always causal and may not mean anything. Founder of the Utrecht Data School (Appendix 8) refers to correlations such as the amount of deaths by drowning and the amount of Nicolas Cage movies, which are not causal but if used can possibly put individuals in groups that they do not belong in.

AWARENESS AND PROTECTION

The interviewees were asked for their vision towards the awareness of the potential risks. It is said that people are generally unaware of what they are giving away when using services. Secondly, people are unaware where their data might end up. Coordinator of a business intelligence center (Appendix 7), describes that there is a small group of citizens that know what has is possible with their personal data, even within the government only a small group of people is aware of the potential. The Chief Privacy Officer (Appendix 3) describes that awareness among citizens is increasing and that organizations are beginning to question whether they even want to get into this type of data analytics. The founder of Utrecht Data School states that individuals do not recognize the issue at hand as it is too complicated for many. Even the decision makers do often fail to ask the relevant questions. As the impact and consequences appear far from people's lives they are not recognizing the issue at hand.

> *It is a danger that people think they are anonymous and that people are unaware of what happens with the data. – Developer of security programs based on behavior – Appendix 5*

> *Not many people have experience with data and until they experience it themselves, it is hard to see the urgency. (…) At the data school we did a lot of work for*

municipalities, councilors and civil servants. What stands out is that they do not ask the right critical questions about digital projects. Projects that include a lot of data. The ownership of the data is often not settled. – Founder Utrecht Data School – Appendix 8

For many people it is no issue at all. It is happening too far away from their lives. For many people it is too complicated. – Founder Utrecht Data School – Appendix 8

The founder of Datafloq explains that a large problem is the awareness of individuals. People do not read terms and conditions when installing applications. He describes that this is partially to blame to the complexity of the terms and conditions but is at large the responsibility of the user.

We are all complaining but we still accept all of it. (…) I am doing it myself, I am very aware of the repercussions to my privacy etcetera but I also disregard the terms and conditions. (…) It is because the terms and conditions are impossible to read. (…) A well-known example is that the terms and conditions of Facebook are more extensive and complex than the American Constitution – Founder Datafloq – Appendix 11

The interviewed speaker on new technologies and biohacker (Appendix 10) describes that it is unclear for people when they can experience the downsides of their data in big data analytics. He continues the explanation by comparing the phenomenon to the climate crisis as for both the consequences are intangible.

You do not know what they can do with my data, I have nothing to hide. If you cannot get a loan from the bank in 20 years, based on data that we are sharing today. People are not likely to know this is the reason. – Speaker on new technologies and Biohacker – Appendix 10.

The interviewed Security Manager and Data Protection Officer (Appendix 4) states that at a previous employer once had to deal with case in which data was leaked. Eventually this ended up in the media but nonetheless of tens of thousands of clients only two asked for their data to be removed entirely. Individuals do not realize that the data they leave behind today can still exist in 30 years. One may be judged in in the future on matters that are shared today (Found Datafloq, Appendix 11).

We are now living in a democracy, but who says that in 30 years we are still living in a democracy? How will our data be treated then? For example, now it is fine to smoke, and everyone is posting picture on Instagram of them holding a cigarette. Maybe in 30 years this is illegal and are we living in a dictatorship and everyone who once smoked is being arrested. A crazy example but it could be a possible reality if we do not treat it in the right way. You can see that in China they are looking different at data and at privacy. I believe that we will stay a democratic country but it is something to consider, data is for a longer period of time. – Founder Datafloq – Appendix 11

Most interviewees mentioned that the GDPR is a step in the right direction when it comes to data- and privacy protection. There is skepticism towards the enforcement of the GDPR, the developer of security programs also states that proper enforcement is not in place. The founder of Datafloq described that regulations always fall behind on technological advancements. He describes that the ethics of the data processors are better protection for the personal data in big data. The coordinator of a business intelligence center (Appendix 7) also states that protection comes at large from the ethics of the people who work with the data. On the contrary, the Professor at Law and Digital society describes that individuals are protected properly by Article 22 of the GDPR and that he sees no reasons to assume that protection is not at the right level (Appendix 6). The Chief Privacy Officer (Appendix 4) states that combinations of data can lead to indirect identification and is thus covered by the GDPR.

I think that by looking at it from the judicial perspective, with the GDPR or AVG we are much better protected [against the downsides]. But the difficulty is in the enforcement. – Speaker new technologies and biohacker – Appendix 10

The law may exist but enforcement does not. – Developer of security programs based on behavior – Appendix 5

(…) the problem is that regulations are always behind on technological advancements. (…) If you care about – Founder Datafloq – Appendix 11

It [protection against downsides] is at large in the ethics of the people working with the data. ICT'ers often have a feeling whether something is allowed from an ethical perspective. – Coordinator business intelligence center – appendix 7

[Combinations of data] make that you are not directly, but indirectly identifiable. When one can be indirectly identified these are personal data and the GDPR is applicable. – Chief Privacy Officer – Appendix 4

The owner of Datafloq (appendix 11) refers to a system that puts the individual in charge of their own data and could offer protection in the future. The system allows individuals to decide for each piece of personal data who gets access to it. Though such as system will not be available in the near future, it is strongly believed in. Another example given that puts the individual in control over their own data is the "self-sovereign identity". The following quote gives an example of the potential of this technology.

If you go to a bar and need to show that you are 18 in order to drink, instead of giving your driver license with all types of personal data, you can cryptographically show that you are 18 or older without showing your actual age and date of birth. Everyone knows the information is correct. It gives your much more control over your data and privacy. – Founder of Datafloq – Appendix 11

With the results found through interviews presented the next chapter uses these results to answer the sub questions of this research defined at the start of this research. Based on this interpretation the hypothesis is confirmed or falsified and the research is concluded.

DISCUSSION

This section attempts to answer the sub questions introduced in the first chapter of this thesis. The results generated through the interviews are interpreted and linked to the literature. By answering the four sub questions it is then determined whether the hypothesis is confirmed or falsified.

CONSEQUENCES IN THE LITERATURE

The first question to be answered is "*What are the potential consequences of personal data processed in big datasets to privacy according to the existing literature?"*. In order to answer this question, the definition of privacy should be taken into account, which is the right to live a life that is private from others without unwanted interference from others, and the right to be in control of access to one's own data. The literature presents three possible consequences of personal data processed in big data. These are re-identification, targeting based on profiles and regarding spurious correlations to make decisions. Firstly, as presented in the literature section, privacy can be affected by the retention of personal data as conclusions can be made on habits, place of residence, daily movements, social relationships and social environments. In order to process these types of data it is often anonymized to hide the identities of data subjects and protect their privacy. In such cases the risk exists that combined sets of anonymized data can be used to re-identify individuals. Various examples are given of cases in which combined anonymized or aggregated datasets could identify the individuals whose data was processed in the dataset. Re-identification is a risk to privacy as it can reveal information about individuals which they did not desire to share. If anonymized datasets are in any way shared with third parties, the third party could potentially identify the individuals and gain access to data that was not intended to be known to them. Considering the definition of privacy, personal data in the hands of unauthorized parties can directly affect privacy as the individual experiences loss of control of access to the data.

Secondly, profiles can be generated by processing personal data in big datasets. Based on these profiles individuals can be categorized, this categorization helps organizations in decision-making and is used to target individuals that possibly fall within a certain profile. While this is often used to benefit the individual by for example giving recommendations on products and services based on their previous behavior, it does bring consequences for the individual. Based on correlations found in combined datasets, the individual can be shown recommendations on products and services that they might like, which they did not want to share with others and may not understand why such recommendations are show. As described by the Dutch expert group on big data and privacy, profiling is considered a risk to privacy as

it can cause individuals to lose control over their freedom to take decisions. Considering the definition of privacy in this research, profiling can affect privacy as profiling can lead to unwanted interference. Next to unwanted interferences, individuals can be excluded or denied access from products or services based on generated categorizations, or profiles. In some cases, the individual may not understand why he or she is being excluded. This can be experienced as an invasion of privacy as individuals experience the sensation of not being able to make their own decision.

The third implication mentioned in the literature is in regards to spurious correlations. Based on corrupt data, discriminatory decision making algorithms can form. Individuals can be wrongfully targeted and excluded based on the conclusions drawn from the biased data. One may not know or understand why they are being treated in a certain way and feel a loss of freedom to make decisions.

To conclude this section and answer the first sub question on the possible consequences of personal data processing in big datasets to privacy it can be said that these consequences are re-identification of individuals based on supposedly anonymous data, targeting based on profiles and interferences based on spurious correlations. Re-identification can result loss of control over data and unauthorized access, targeting can result in unwanted interferences and exclusion can be experiences as loss of freedom to make decisions. The next section will examine the definition of personal data as given by the interviewees.

PERSONAL DATA

The second question is "*Is personal data perceived differently between interviewees?".* As presented in the chapter containing the interview results, the interviewees generally agreed that all data that can be traced back to the individual is in fact personal. When asked what can be considered personal data the interviewees gave a range of answers, one more far reaching than others. While some interviewees mentioned that patterns of online clicking behavior can be classified as personal data others disagreed as this data could not always directly identify an individual. One interviewee insisted that details on whether one locks their computer at work are personal data and another that a combination of only search queries can be personal data. On the latter, the literature does agree as a so-called clickstream linked to an individual becomes difficult to detach from an individual. The interviewees see anonymized personal data as personal data due to the re-identification that can take place by linking datasets or through unique combinations. The definition of personal data did largely match the extensive definition in de GDPR. Even though all interviewees perceive data that directly identifies an individual as personal data, some have further reaching definitions. The answer to the second sub question of this thesis is yes, personal data is perceived differently

between interviewees. For the remaining results the personal data will be all data that can directly or indirectly identify an individual.

As the interpretation of the concept personal data differs between experts it can be wondered how different organizations interpret it. Organizations may be processing personal data but not recognize this as such, which could result in less extensive protection. As mentioned during the interviews, the definition is being broadened continuously. The description of privacy by Warren and Brandeis, as mentioned in the conceptualization, already stated that what is to be protected as privacy needs to be revised from time to time. With increasing amounts of personal data being generated the need for revision may be necessary more quickly than in times when less data was generated.

UP- AND DOWNSIDES

The third question is "*What are the up- and downsides of personal data in big datasets for individuals?"*. Based on the literature and interviews three upsides are identified, these are personalized content, increased ease of use and better products which is very close to personalized content. The identified downsides are four in total. These are exclusion, no control over choices, no control over data and finally the limitation of information. The following paragraphs will explore both the upsides and downsides. First the upsides will be explored and secondly the downsides.

The first and biggest upside is the fact that based on profiles, which are created based on their own behavior and the behavior of similar people, the individuals in big datasets will see information relevant for their profile. Based on the personal data in big data the individual can be offered the right product, through the right channel at the right time. The individual can be given better treatment or discounts on products and services because they fit a certain profile. If data in the set used to give recommendations is 100% accurate, the individual would never have to see irrelevant content again. The second upside, next to content or pricing based on profiles, the ease of use increases due to personal data processed in big data. Large amounts of data on usage allows service providers to optimize their product. For example, while using Google services, it automatically suggests recipients of e-mails, finishes your sentences or suggests websites based on previous behavior. Facebook does the same by suggesting friends or showing events that you might like. Suggestions while using services increase the ease of use as the user has to give less inputs to reach their desired goal. Lastly the individual can be offered a better product as products can be made close to perfect based on data. Data on malfunctions in the product itself can be used in analyses and help the manufacturer to make their product better. An example of this is when a manufacturer of printers analyzes the product data and notices that a certain part is likely to wear out much

quicker than the rest of the machine. The manufacturer can decide the make implement a sturdier part to make the machine more durable. In this example the data comes from the hardware and is based on usage behavior, which can arguably be seen as personal data.

A total of four downsides for individuals are mentioned in the results of this research. These are exclusion, no control over choices, no control over data and finally the limitation of information. The first downside for individuals whose data is being processed is exclusion. Based on the previously described profiles that can benefit individuals, they can also be excluded. During the interviews it has frequently been mentioned that big data analyses can exclude individuals and could form discriminatory decision making models. Products and services that individuals may previously have enjoyed access to can decline them based on insights gathered through big data analysis. If the data in the analysis is seen as accurate the bias which creates the discriminatory model can possibly reproduce and reach even further. By relying on such models the same groups can continuously be excluded. In some cases, the individuals may not know why they are being refused.

This lays closely to the second downside for individuals, due to a lack of transparency in the decision making, or to algorithms used to make decisions, one may feel loss of freedom over their own choices. Big data analysis can draw conclusions based on many combinations of data from different sources. The conclusions based on this data may not be comprehensible for individuals that are being categorized, even if they are given access to the rationale behind it. Rather than basing decisions on traditional data, such as data on income to determine whether someone is eligible for a loan, decisions can be based on algorithms. These big data powered algorithms can take into account other types of personal data and come to a different conclusion. Without transparency individuals do not know why decisions are made and may feel that they are not to make their own decisions.

The third downside for individuals is the loss of control over personal data. While companies must explain their purpose for the data collection upon asking consent, this does not give the individual any guarantees. When the data is provided to a data processor it is no longer in control of the individual. Data could be anonymized sufficiently at this point in time but as more data is created, an anonymized dataset combined with another can potentially re-identify individuals in the future. The individuals whose data is being processed have no guarantee that the data will only be used for the consented purpose. The individuals providing the data are protected against this by the GDPR but doubt is expressed towards the enforcement of this right. When allowing organizations to process personal data, individuals are to trust the ethics of the processors as it can hardly be controlled what is actually done with the data. The fact that personal data can end up in unauthorized hands puts privacy at risk. The mentioned database of doom in the wrong hands can have serious consequences for one's private life.

The fourth downside is the limitation of information. While individuals are enjoying personalized content based on determined profiles they are being only shown information that is close to their personal opinion and interest. This can result in vast quantities of available information not reaching individuals. These persons may lose perspective and can be influenced as they are possibly only shown information that they agree with. This is close the second downside mentioned above as individuals may for example be influenced in their voting behavior and unknowingly lose control over this choice.

To answer the third sub question, the upsides of personal data processed in big datasets for individuals are personalized content, increased ease of use and a better product. The downsides are possible exclusion, loss of control over choices, no control over personal data and limitation of information.

POSSIBLE EFFECTS TO PRIVACY

The fourth question of this thesis is close to the main research question and the first sub question. It attempts to identify the possible effects of personal data processed in big datasets to privacy based on the literature and interviews combined. The results from the literature are close to the results gathered through the interviews. Four effects were found and are explained in this section, these are unwanted interferences, persuasive pressures, loss of control over data and exclusion.

First unwanted inferences, nearly every service and application that people use in daily life collects data and people tend to agree to nearly all of it. As mentioned during the interviews, individuals are easy to give consent to data collection without considering the possible consequences. Individuals can be exposed to content or products based on profiles generated through behavior, that they would not want to have shared with anyone. If the interference is precisely on what the user has given consent to it would no longer be unwanted and privacy remains unaffected. If organizations that process data, collect data with the purpose to target individuals, this targeting does not affect their privacy as the individual has given authorization. If organizations collect data for a different purpose and then uses this data to make interferences, it is unauthorized and affects privacy. The privacy norm as described in the conceptualization of privacy is important to consider here. As presented in the research conducted by the Eurobarometer, 58% of participants do not see an alternative to sharing their personal data if they want to use online products and services. In other words, for more than half of the participants the norm is to share personal information as they see no alternative. While more than half of the participants does not see an alternative, 60% does not like giving up the information in return for free services. The created norm is giving consent to personal data processing while the majority of the participants does not like giving up this

information. During the interviews it was mentioned that individuals simply agree to terms and conditions as they are too complex and that individuals should take responsibility in this matter. It can be argued that due to created privacy norms individuals do not protect their own privacy while research shows that people are worried about it. Without thoroughly considering the potential implications individuals share personal data. Based on this data profiles are made that are used to target individuals. This targeting can be in the form of unwanted interferences.

Secondly, based on the personal data that is processed, individuals can be shown content that can influence them, especially if they are only shown content that may push them in a certain direction. While individuals may enjoy having their newsfeed filled with articles of their own interest, it does keep information away from them. The example of providing different types of information during election campaigns is an example of this. As individuals from a certain political party are only shown information that matches their views, they can be influenced to vote for this party again. Information about the party that matches less with their interest can be kept away from the individual while this could possibly cause the individual to vote differently. This is affects the psychological dimension of privacy as individuals can be subjected to persuasive pressures and be influenced in their thoughts, feelings, attitudes and values.

In the big data context it is challenging to use data for just one purpose as big data's strength is in the combination of different datasets. As mentioned during the interviews big data and privacy are at odds with each other. This is due to the fact that big data analysis can use varying data of large quantities, if even the slightest amount of personal information remains it can possibly be used to identify an individual. Organizations may want to keep datasets available for various analyses. In order to keep data, it can be anonymized, which supposedly removes all personal identifiers. This brings the third effect, which is the loss of control over one's own data. Anonymized data that cannot directly or indirectly identify individuals is not regarded as personal data by the GDPR, and can be used for other purposes or even be shared with other organizations. As became clear during this research, anonymization is not regarded as a valid strategy and re-identification of individuals is likely. The data provided by the individual is no longer within their control and it can possibly be shared in anonymous formats. As the data can be de-anonymized and used to identify individuals they experience a loss of privacy. Organizations can fail to see the significance of personal data and be reluctant towards the ownership of the data. Private information that was only intended to be shared with one organization can end up at organizations with very different goals. This is an invasion of the information dimension of privacy as the individual loses the ability to control who gathers and spreads information. The privacy norm, or

purpose of the data processing, may be violated as the usage of information can be different than originally consented on.

The fourth effect also refers to unwanted interferences. The interviewees and literature expressed doubt towards the quality of data in big data analyses. If due to any reason, the data is not correct and used to come to conclusions on individuals they may be targeted with information and offers that are entirely irrelevant. The individuals can be excluded based on the near discriminatory profiles created with biased algorithms. Interviewees expressed great worry towards individuals being excluded and not understanding why they were being excluded. This can be experiences as a loss of freedom to make decisions. This exclusion is an invasion of the psychological dimension as individuals are categorized and nearly unable to fight this pressure.

Considering the four effect that have been explained it can be established that privacy, defined as the right to live a life free from others, without unwanted interference and to be in control over one's own data, can be affected by personal data processed in big data. The hypothesis, *"Privacy can be affected by processing personal data in big datasets",* which was central throughout this research, is thus confirmed. Notable is that according to the interviewees, the individuals whose data is being processed are not aware of the potential effects to privacy. People are used to the comforts of personalized content and are said to be barely interested in the downsides. Even people that take decisions on big data processing are considered not aware of big data's potential and the possible effects to privacy the processing can have. The Eurobarometer shows that only 9% of the participants feels that they are completely in control over their data. The low percentage of total control explains why individuals worry about their privacy, especially if 58% of the participants feel that they have to give up personal information in order to use the services they want. Individuals feel forced to give up information to organizations such as social networks and search engines while only 18% trusts these organizations. The low degree of trust and low feeling control explains why individuals worry about their privacy. Research by *Autoriteit Persoonsgegevens* shows that 94% of individuals worries about their privacy while 88% has never used their right to privacy. This matches with what has been said during the interviews as individuals can hardly grasp the potential impact of personal data processing to their privacy. It is seen as too complex and not many have felt the direct consequences before. This could be a reason why 88% has never acted on their privacy.

CONCLUSION AND RECOMMENDATIONS

To main question of this research is: *"How can privacy be affected by processing personal data in big datasets?"*. The answer to this question is as follows, the processing of personal data in big datasets can affect privacy as it can result in unwanted interferences, persuasive pressures due to information limitation, loss of control over data and exclusion. It should be considered that individuals often consent to what data processors can do with their data. In regards to unwanted interferences it should be noted that if consent to data processing is given, and the data processor does only what is given consent to, it is not affecting privacy of individuals as the interference is authorized. When organizations do anything with the data that the individual has not given consent to it can result in unwanted interferences and thus affect privacy. Even though persuasive pressures are allowed as the individual has given consent to personalized content and is thus only shown information close to their own interests, it can affect the psychological dimension of privacy. It is difficult for individuals to know whether organizations are only doing what they have consented to. Individuals do not see other options than to share their personal data while the majority does not enjoy giving up this information. Nonetheless, the processing is often given consent to. In some cases data can be anonymized in order to be allowed to use it, as the usage of data may not be allowed with personal identifiers present. Anonymized data can be shared with other organizations while the research has shown that anonymized datasets can likely be used to re-identify individuals. This can result in loss of control over data for the individual and affect their privacy. The final way through which privacy can be affected is by exclusion based on spurious correlations found in big datasets. The near discriminatory algorithms on which decisions are based can exclude individuals whom may feel loss in freedom to take decisions.

The body of knowledge used for this research has allowed for answering of the main research question. The results from the literature and the interviews have common ground and allowed for a grounded answer to the main research question. Total saturation in interview responses has not been reached, which means that there could be possible consequences of personal data processed in big datasets that were overlooked.

The conducted research strengthens academic knowledge as the field has been explored through a qualitative approach. Most available research has been done using statistical approaches and experiments that attempt to de-anonymize data. To further enrich academic knowledge in this field it is recommended that future research is conducted on the following topics. It can be doubted whether current technology is capable of properly protecting privacy. During the interviews it was mentioned in multiple ways that big data and privacy are at odds with each other. Research is to be conducted that examines the potential of new technologies that put the individual in control of their own data. The second topic on

which future research is recommended is whether people would be more careful to consent to personal data processing if they knew what giving up the data can practically mean in the future. It has been mentioned by experts that people are reluctant towards their data protection because they do not understand the matter at hand. As data left behind today can possibly have consequences for the future this research could create more awareness in society.

It is recommended that individuals take responsibility in the protection of their own personal data. Consent is easily given without considering the consequences. This is partly to blame to the reluctant attitude of individuals but also to the organizations that process data. Their terms and conditions that are to be agreed to are often very complex and uninviting to read. Organizations are to be more transparent in the ways that personal data is used and present this to the users in a more readable manner.

BIBLIOGRAPHY

Anthony, D., Campos-Castillo, C., & Horne, C. (2017). Toward a Sociology of Privacy. *Annual Review of Sociology*, 249-269.

Autoriteit Persoonsgegevens. (2018). *Richtsnoeren inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verodening (EU) 2016/679.* Den Haag.

Autoriteit Persoonsgegevens. (2019). *Nederland maakt zich zorgen over privacy.* Den Haag: Autoriteit Persoonsgegevens.

Autoriteit persoonsgegevens. (n.d.). *Wat zijn persoonsgegevens?* Retrieved from Autoriteit Persoonsgegevens: https://autoriteitpersoonsgegevens.nl/nl/over-privacy/persoonsgegevens/wat-zijn-persoonsgegevens

Barbaro, M., & Zeller Jr, T. (2006, 08 09). *A Face Is Exposed for AOL Searcher No. 4417749*. Retrieved from The New York Times: https://www.nytimes.com/2006/08/09/technology/09aol.html

Bredenoord, A. L., van Delden, J. J., Mostert, M., & van der Sloot, B. (2017). From Privacy to Data Protection in the EU: Implications for Big Data Health Research. *european Journal of health law 24*, 1-13.

Burgoon, J. K., Parrott, R., Le Poire, B. A., Kelley, D. L., Walther, J. B., & & Perry, D. (1989). Maintaining and restoring privacy through communication in different types of relationships. . *Journal of social and personal relationships*, 131-158.

Byford, S. (1996). Privacy in cyberspace: Constructing a model of privacy for the electronic communications environment. *Rutgers Computer and Technology Law Journal*, 1-74.

Cadwalladr, C., & Graham-Harrison, E. (2018). The Cambridge analytica files. 6-7. Retrieved from The Guardian.

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science*, 595-612.

Cavoukian, A., & El Emam, K. (2011). *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy.* Canada: Information and Privacy Commissioner of Ontario.

Council of the European Union. (2015). *The General Data Protection Regulation.*

Crovitz, L. (2012, November 18). *Obama's 'Big Data' Victory.* Retrieved from Wall Street Journal: https://www.wsj.com/articles/SB10001424127887323353204578126671124151266

Datoo, A. (2017, 07 24). *GDPR and Big Data - Friends or Foes?* Retrieved from GDPR:Report: https://gdpr.report/news/2017/07/24/gdpr-big-data-friends-foes/

Dwork, C. (2008). An Ad Omnia Approach to Defining and Achieving Private Data Analysis. *International Workshop on Privacy, Security, and Trust in KDD*, 1-13.

European Commission. (2015). *Eurobarometer Gegevensbescherming.* European Commission. Retrieved from http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_fact_nl_nl.pdf

Expertgroep Big data en privacy. (2016). *Licht op de digitale schaduw verantwoord innoveren met big data.* Den Haag: Ministerie van Economische Zaken.

GDPR Report. (2017, 07 24). *GDPR and Big Data - Friends or Foes?* Retrieved from GPDR:REPORT: https://gdpr.report/news/2017/07/24/gdpr-big-data-friends-foes/

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *British Dental Journal*, 291-295.

Gonzalez, R. (2017). *Hacking the citizenry?: Personality profiling, 'big data'and the election of Donals Trump.* 9-12: Anthropology Today.

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013). A discussion of privacy challenges in user profiling with big data techniques: The EEXCESS use case. *2013 IEEE International Congress on Big Data*, 25-30.

Irani, D., Webb, S., Li, K., & & Pu, C. (2009). Large online social footprints--an emerging threat. *2009 International Conference on Computational Science and Engineering*, 271-276.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *IEEE Computer Society*, 56-59.

Jensen, M. (2013). Challenges of privacy protection in big data analytics. *IEEE International Congress on Big Data*, 235-238.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences.* London: Sage.

Mathers, N., Fox, J., & Hunn, A. (1998). Using interviews in a research project. *Research Approaches in Primary Care*, 113-134.

Mayer-Schönberger, V. &. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.

n.a. (2019). *Focused (Semi-structured) Interviews.* Retrieved 05 15, 2019, from http://www.sociology.org.uk/notes/methfi.pdf

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 ieee symposium on security and privacy*, 111-125.

Netflix. (n.d.). *The Netflix Prize Rules*. Retrieved 04 01, 2019, from Netflix Prize: https://www.netflixprize.com/rules.html

Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA*, 1701-1777.

Polonetsky, J., & Tene, O. (2013). Privacy and Big Data: Making Ends Meet. *Stan. L. Rev. Online 66*, 1-7.

Schermer, B., Hagenauw, D., & Falot, N. (2018). *Handleiding Algemene verordening gegevensbescherming en Uitvoeringswet Algemene verordening gegevensbescherming.* Den Haag: Ministerie van Justitie en Veiligheid.

Statista. (n.d.). *Number of monthly active Facebook users worldwide as of 4th quarter 2018 (in millions)*. Retrieved 02 27, 2019, from Statista: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Supriyadi, D. (2017). *Personal and Non-Personal Data in the Context of Big Data.* Tilburg: Tilburg Institute for Law, Technology and Society.

Tene, O., & Polenetsky, J. (2011). Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, 63-69.

Tene, O., & Polonetsky, J. (2013). Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property*, 239-273.

Torra, V., & Navarro-Arribas, G. (2016). Big data privacy and anonymization. *IFIP International Summer School on Privacy and Identity Management*, 15-26.

Vedder, A. (2009). Privacy, een conceptuele articulatie. *Filosofie & praktijk*, 7-19.

Vetzo, M., Gerards, J., & Nehmelman, R. (2018). *Algoritmes en grondrechten.* Den Haag: Boom Juridisch.

Warren, S., & Brandeis, L. (1890). The Right to Privacy. *Harvard Law Review*, 193-220.

Westin, A. (1967). *Privacy and freedom.* New York: Atheneum.

Wetenschappelijke Raad voor het Regeringsbeleid. (2016). *Big Data in een vrije en veilige samenleving.* Den Haag: Rapport 95 .

White House. (2014). *Big Data. Seizing opportunities, preserving values.* Washington D.C.: Executive Office of the President.

APPENDIX 1: INTERVIEW QUESTIONS

Algemene vragen

Zou u uzelf kort willen voorstellen?

Kunt u uw vaste werkzaamheden omschrijven?

Ervaring met big data

Kunt u beschrijven wat big data is?

Zijn er bepaalde kenmerken voor data om als big data gezien te worden?

Op welke manier heeft u ervaring met big data?

Voor welk doeleinde wordt big data naar uw weten gebruikt?

Hoe wordt big data binnen uw bedrijf gebruikt?*

Kennis over persoonlijke data in big datasets

Wat is persoonlijke data?

Hoe komt, in uw optiek, persoonlijke data terug in big datasets?

Is persoonlijke data cruciaal in big datasets?

Hoe wordt persoonlijke data verwerkt in big datasets binnen uw bedrijf?*

Voor- en nadelen van persoonlijke data in big datasets

Wat zijn de voordelen voor personen van hun data in big datasets?

Wat zijn de nadelen voor personen van hun data in big datasets?

Is er voldoende bescherming tegen de mogelijke risico's?

Effecten persoonlijke data in big datasets

Kunt u beschrijven wat privacy is?

Is anonimiseren van persoonlijke data een methode waardoor persoonlijke data wel gebruikt kan worden in big data?

Eerdere onderzoeken tonen aan dat data in big datasets geanonimiseerd dient te worden zodat het niet direct herleid kan worden naar een individu. Denkt u dat deanonimisering een reëel risico is voor privacy van individuen en kunt u dit toelichten?

Hoe zouden de door u genoemde effecten het best aangepakt kunnen worden?

Denkt u dat er genoeg aandacht is voor de risico's van persoonlijke data in big datasets?

Denkt u dat deze risico's in de toekomst groter worden?

Afsluiting

Denkt u dat er aspecten zijn van dit onderwerp dat ik over het hoofd zie waar ik ook nog naar zou kunnen kijken?

*= *Indien toepasselijk*

APPENDIX 2: LEGEND INTERVIEW CODING

Theme	**Color coding**
Privacy	Bright yellow
Big data (description, goal, practices)	Bright green
Personal data (description, part of big data)	Bright blue
Anonymizing	Pink
Upsides	Teal
Downsides	Dark red
Protection against risks	Violet
Attention for risks	Dark yellow
Profiling (categorizing)	Bright red
Spurious correlations (quality of the data)	Grey
Future	Green