Master's Thesis

Examining the *post-privacy* world

R.E.Y. Haasjes

First supervisor:
Dr. Kevin Macnish

Second reader:
Dr. Brandt van der Gaast

Philosophy of Science, Technology and Society (PSTS)
University of Twente, Faculty of Behavioural, Management, and Social Sciences
Enschede, the Netherlands, *October 2018*

Table of Contents

Summary

In a recent paper, Wang and Kosinski claim that machine learning classifiers can accurately detect sexual orientation from facial images. They believe that due to the growing digitalisation of our lives and the rapid progress in artificial intelligence, we are inevitably headed towards a world in which privacy has been completely eroded, what they call “the post-privacy world”. This thesis examines this post-privacy narrative, by questioning whether predictions by machine learning classifiers can violate one’s privacy, if we assume that the access account of privacy is correct. By assuming the access account of privacy, this thesis focusses specifically on what information is accessed, and what is uncovered by machine learning classifier predictions.

This thesis shows that predictions by machine learning classifiers could potentially violate privacy. First of all, in order to make the predictions, machine learning classifiers have to be trained, which is often done using data that is taken out of context, breaching contextual integrity and privacy in the process. In addition, the existence of a machine learning classifier that could uncover private information from public information does not take away the reasonability of a claim to privacy with respect to this information. Last, but definitely not least, due to the technological sophistication and the vast amount of data used in machine learning, predictions by machine learning classifiers have acquired an unjust amount of epistemic status. Due to this unjust epistemic status, the danger exists that predictions by machine learning classifiers are assumed to be privacy invasive, even when there is no strong evidence that they are. These cases in which privacy is not violated, could be just as harmful, perhaps even more harmful, as when privacy is violated.

Acknowledgements

First and foremost, I want to thank my first supervisor, Dr. Kevin Macnish, for his patient guidance and enthusiastic encouragement. He provided valuable insights, suggestions and directions in our meetings, and engaged in various silly thought experiments. I am also grateful to my second reader, Dr. Brandt van der Gaast, for his useful and insightful comments. Lastly, I want to thank my family and friends, especially my girlfriend and father, for their loving support and providence of food throughout my studies.

Introduction

In a recent paper, Yilun Wang and Michal Kosinski claim that machine learning classifiers can accurately detect sexual orientation from facial images (Wang & Kosinski, 2018). They used a deep neural network to extract features from 35.326 facial images. These features were used as independent variables in a logistic regression classifier, with self-reported sexuality being the dependent variable. Wang and Kosinski report that this trained logistic classifier, given a single image, could correctly distinguish between a gay and heterosexual man in 81% of cases, and in 74% of the cases for women. If the classifier was given five images, this percentage increased to 91% for men and 83% for women (Wang & Kosinski, 2018, p. 250). Furthermore, they conducted a number of studies from which they concluded the following (Wang & Kosinski, 2018, p. 254):

- Study 1b: the predictions were based on the part of the image that contained the face and not on the background
- Study 1c: gay man and lesbian woman had gender-atypical features
- Study 2: the probability of being gay was positively correlated with facial femininity among males and negatively with female facial femininity
- Study 3: a lot of information about sexual orientation is retained in fixed facial features
- Study 4: non standardized facial images were not especially revealing of sexual orientation
- Study 5: the classifier performed similarly with facial images collected in a different environment

Based on these results, Wang and Kosinski conclude that facial images contain more information about sexual orientation than the human brain can perceive and interpret (Wang & Kosinski, 2018, p. 254). Wang and Kosinski elaborately stress the importance of the ethical implications of their research. They discuss how previous research has shown that intimate information such as one's sexual orientation can be revealed by Facebook likes (Wang & Kosinski, 2018, p. 255). In addition, similar research aimed to show that Facebook friendships can expose sexual orientation (Jernigan & Mistree, 2009). However, whereas Facebook likes and other digital footprints can be hidden and anonymized, it becomes increasingly difficult to hide one's facial information. Wang and Kosinski believe that the accuracies reported in their study are also not the upper bound of what is possible: they used widely available off-the-shelf tools, publicly available data and well-known methods. With more information and more sophisticated techniques, accuracy could potentially be increased (Wang & Kosinski, 2018, p. 255). Due to the accuracies reported in their studies, and the potential for even higher accuracies, they believe that the growing

digitalization of our lives and rapid progress in artificial intelligence will erode the privacy of intimate traits such as sexual orientation (Wang & Kosinski, 2018, p. 255). They label the world in which privacy has been completely eroded the "post-privacy" world (Wang & Kosinski, 2018, p. 256).

Wang and Kosinski fear for the safety of gay people in countries and cultures in which homosexuality is not accepted. They state that some governments are already developing and using face-recognition software with the aim of detecting intimate traits, making the need for awareness of these technologies among homosexual communities, policy-makers and governments even more urgent (Wang & Kosinski, 2018, p. 255).

Defining privacy

Although the ethical concerns posed by Wang and Kosinski seem realistic and well-meant, assuming that we are inevitably headed towards a *post-privacy* world severely limits the extent of their ethical examination. First of all, there is an ongoing discussion on how privacy should be defined. Kevin Macnish contributes to the debate about two major definitions of privacy: the control and the access accounts (Macnish, 2016). Macnish argues that the access account is correct and the control account is mistaken. He argues this mainly through a thought experiment in which a person loses their diary. If a person loses their diary in a coffee shop and later find a stranger holding their diary, it might feel like an invasion of privacy, but Macnish argues that the person's privacy is only invaded when the diary is actually read by the stranger. However, Macnish does not argue that seizing control over someone's personal information is therefore unproblematic. He argues that seizing control over someone's information can be harmful, in some cases even more harmful than violating someone's privacy (Macnish gives the example of being blackmailed by a stranger who holds control over your diary, but who has in fact, not read your diary).

To prevent the risk of making this thesis about which definition of privacy is correct, instead of on the ethical implications of machine learning on privacy, I will assume that the *access account of privacy* is correct. It seems reasonable to question whether the definition of privacy is important with respect to examining the ethical implications of machine learning classifiers on privacy. I argue that there is something valuable in making the distinction between the control and the access account of privacy.

Assuming the access account of privacy to be correct allows us to focus on the information that is actually generated, or uncovered by machine learning classifiers. It seems obvious that control over one's private information can be lost due to these machines, and consequently, this can be harmful. If a totalitarian state which is strictly against homosexuality considers 91% accuracy to

be sufficient, the people in that state would lose control over their ability to decide to whom they want to reveal their sexual preference. However, the factual private information could potentially stay private. Nevertheless, these people could be severely harmed, put into prison or even executed. Therefore, it is of vital importance to examine what information these machine learning classifiers actually *access* or uncover.

Outline

The main research question in this thesis is: *"Can predictions by machine learning classifiers violate one's privacy?"* At first glance one might think this question has an obvious answer, namely that predictions of machine learning classifiers can violate one's privacy, if private information is inferred from non private information. However, throughout this thesis I will show that the answer to this question is not as straightforward as it seems, given that we assume the access account of privacy as correct. In the first chapter, I place machine learning in the broader context of the practice of data mining, and clarify the terminology used throughout this thesis.

In chapter two, contextual accounts of the privacy implications of data mining are analysed. Using the contextual accounts of privacy one of the problems at the heart of data mining is identified, that in the process of data mining, data is often taken from one context and used in another. However, data could be public in one context, but at the same time, deeply private in another. Whereas chapter two focusses on privacy issues related to data that is used in a different context, chapter three focusses on the nature of the machine learning classifier predictions. If machine learning classifiers are able to accurately uncover private information from non private information, one could argue that by sharing the non private information, one gives up the right to privacy with respect to the private information. I will tackle this problem using Jason Millar's concepts of core private information and privacy, and show that even though it could be possible to uncover private information from information that was willingly shared, this does not entail that one cannot make any post hoc claims to privacy with respect to the uncovered information.

In chapter four I criticise some of the assumptions that Millar makes in his analysis of the privacy issues raised by predictive mining. I will show that in order to give an adequate answer to the question of whether predictions by machine learning classifiers can violate one's privacy, we should look at the broader context of the machine learning classifier to determine whether the predictions are credible. In chapter five I elaborate on this, by examining in more detail how we should determine the credibility of machine learning predictions. This is done by viewing privacy violations according to the access account of privacy as learning something private about an individual. I will show that although it is possible for machine learning classifiers to uncover

information that is informative about an individual and potentially privacy invasive, it might be more problematic when the predictions are not privacy invasive, but are considered privacy invasive due to their elevated epistemic status. Lastly, I will conclude and discuss some limitations and future research suggestions.

Chapter 1

Data mining and terminology

In this chapter, I will briefly discuss concepts in the practice of data mining and clarify terminology that will be used throughout this thesis. Wang and Kosinski make a reasonable point by arguing that anonymising people's faces would be difficult in practice. If Wang and Kosinski's claim that sexual orientation can be accurately detected from facial images alone turns out to be true, this could have severe consequences for people's perception on the private / public status of personal information surrounding one's sexuality. Is it possible that we are heading towards a world in which everyone wears a burqa to protect their face from being recorded, or will people's sexuality become public information? Nowadays, people are willing to share their faces on the internet. Instagram, Facebook and Twitter are all platforms on which many people love to share photos of themselves and loved ones. It is this willingness to share information and the public nature of the internet that gave rise to the growing practice of data analytics. As early as 1985, Larry Hunter, a computer scientist observed the following: *"Our revolution will not be in gathering data – don't look for TV cameras in your bedroom – but in analysing the information that is already willingly shared"* (as cited by Nissenbaum, 1998, p. 560).

The practice of trying to uncover new meaningful information from data has been around for a long time. In the past, this practice was often described as Knowledge Discovery in Databases (KDD). Herman Tavani describes KDD as the overall process of discovering useful information from data, which includes gathering data, processing data, mining data and interpreting this data (Tavani, 1999b). Data mining is one of the most discussed steps in this process, which combines artificial intelligence, statistical analysis, knowledge acquisition from expert systems, data visualization, machine discovery and pattern recognition (Tavani, 1999a). KDD has come a long way since then. Whereas in the past primarily numerical, and structured data stored in traditional databases was analysed, nowadays heterogeneous data sources are analysed which include structured, semi-structured and unstructured data (Venkatadri & Reddy, 2011). Due to the massive use of the internet through the last decades and the rise of social media, massive amounts of data are currently available which has led to a so called "*big data era*". Therefore, nowadays we often speak of big data analytics, which aims to uncover useful information from various heterogeneous data sources. Furthermore, due to the increasing availability of computing power and improvements in

data mining technologies, data analytics have become more accessible and sophisticated. Machine learning is one of the techniques that is used in data mining. Currently, one of the most popular research fields in machine learning is deep learning (Qiu, Wu, Ding, Xu, & Feng, 2016), which was the machine learning technique used by Wang and Kosinski. Qiu et al describe deep learning as a technique that uses mathematical models which are inspired by the human brain to automatically learn data representations, from large volumes of raw data.

Since the running example in this thesis is the homosexuality classifier of Wang and Kosinski, which uses machine learning, specifically deep learning, I will use the term *classifications* of machine learning classifiers interchangeably with *predictions* by data mining. Although I will sometimes briefly mention potential issues with the practice of data mining that focusses on gathering information from various sources to create profiles of individuals, I will focus on data mining that uses machine learning in an attempt to uncover patterns and underlying structures in large sets of data.

Chapter 2

The privacy implications of data mining

In the previous chapter a brief overview of the practice of data mining was given. Although the type and amounts of data collected and analysed nowadays differ from the past, as early as 1998, people were examining the ethical implications of data mining, specifically related to privacy (Nissenbaum, 1998; Tavani, 1999a). In this chapter, I will examine existing analyses of data mining and privacy and critically evaluate their relevance and importance with respect to contemporary data mining practices, taking the homosexuality classifier of Wang and Kosinski as a running example. I will build on the work of Jason Millar, who wrote an article on privacy concerns raised by predictive data mining (Millar, 2009). Millar is fairly succinct in his analysis of existing work, but he briefly discusses the work of Herman Tavani and Helen Nissenbaum on privacy, data mining and information technology. I will examine the works of Tavani and Nissenbaum in more detail, and at the same time critically examine the critiques and comments of Millar.

2.1. Contextual privacy

Millar observed that both Nissenbaum and Tavani analyse the privacy implications of data mining from the perspective of the contextual aspect of data (Millar, 2009, p. 110). In this section, I will briefly summarise this contextual analysis of the privacy implications of data mining. Tavani illustrates the potential privacy implications of data mining using a hypothetical case of Lee, a junior executive at the ABC Marketing Firm in the United States, who applies for an auto-mobile loan at a local bank (Tavani, 1999a, pp. 140-142). To acquire this loan, Lee gave the bank personal information, such as information about his employment, his salary and savings. Giving the bank access to this information seems reasonable, since the bank needs appropriate information to make a decision on whether they will grant Lee an auto-mobile loan or not. Tavani continues this hypothetical case by stating that the bank then uses data mining techniques, using Lee's and other individuals' data, to find out that people with similar earnings, properties and employment often start their own business within five years, and often declare bankruptcy within one year of starting this business. Tavani argues that the data mining in this example is problematic, because even though individually, each piece of personal information was voluntarily given to the bank, each piece of information was given for a specific purpose and use, determining whether Lee could get a

loan or not. However, by no means did Lee authorise the bank to use the individual pieces of personal information for more general data mining purposes (Tavani, 1999a, p. 142).

Millar discusses how the argument of Tavani depends on a *contextual account of privacy*, which rejects the traditional public-private dichotomy of information, and instead, holds that information can be private in one context and public in another (Millar, 2009, p. 110). Millar illustrates this using the example of a person walking down Main Street in the heart of a local gay district. The knowledge that the person was walking there could be public with respect to certain friends of the person, but deeply private with respect to the co-workers of the person (Millar, 2009, p. 108). This contextual account of privacy and the access account of privacy can be held as true simultaneously. One can allow certain people in a specific context access to personal information, but consider it a severe privacy violation if the same information was accessed by other people or in a different context. But a violation of privacy has only taken place when information was actually accessed.

Turning back to the example of Lee, Millar argues that the data mining is problematic because it shift the context in which the data of Lee is considered (Millar, 2009, p. 109). According to Millar, Lee grants the bank access to his personal information for his request for a loan, but does not grant the bank access to the data for more general analyses and predictions on his future credit risk. Although I agree with Millar that the data mining is problematic due to the shift in context, the problem does not arise because of the attempts to predict the future credit risk of Lee. Given the context of applying for a loan, it seems reasonable to make an estimate of how likely Lee is to pay back the loan or not. Tavani primarily stresses that by using Lee's data combined with other individuals' data for more general data mining, the bank used information about Lee in a way that Lee had not explicitly authorized (Tavani, 1999a, p. 141). Because the bank used the data of Lee outside of the original context, determining whether Lee could get a loan or not. Instead, they used his and many other peoples data with the aim of discovering patterns in the aggregated dataset, which goes beyond what Lee initially gave them the data for. Using this example, Tavani concludes that data mining is clearly incompatible with two principles as specified in the Code of Fair Information Practices: purpose specification and use limitation (Tavani, 1999a, p. 142). The two principles describe that the purpose of which data is collected should be specified at the time of data collection and the data should not be used or made available in ways different than specified in the purpose specification (Tavani, 1999a, p. 142). Going back to Lee, Tavani describes how the data was used in different ways than Lee had consented to in the purpose specification.

2.1.1. The knock-down argument

Tavani's analysis gives insight in why a lot of data mining practices feel like an invasion of privacy: because information that was given in a specific context is used in a different context. However, Tavani's analysis does not give a full account of how the homosexuality classifier of Wang and Kosinski should be evaluated. In the example of Lee, personal information was only given to the bank, stored in an internal database for data mining. However, Wang and Kosinski scraped the data to train their classifier from a dating website, on which people willingly revealed information about themselves to other people on the dating site. Furthermore, their classifier could be applied to classify the sexuality of individuals who have posted photos of their face on personal websites, blogs, social media and other places on the internet. Tavani also observed this distinction, but at the time that he wrote his article data mining from websites was not as common as nowadays (Tavani, 1999a, p. 143). Tavani did predict the potential issues with the internet as a potential data mining resource: *"However, what distinguishes the Internet as a potential mining resource from large commercial databased used in data mining is the vast amount of non transactional, personal information currently available on the Web that could also be mined. Can this personal information, which is also public in some sense, be protected?"* (Tavani, 1999a, p. 144)

Millar states that both Nissenbaum and Tavani acknowledge that a contextual account of privacy has to deal with a normative *"knock-down"* objection, *"that profiling is acceptable because the individuals whose privacy is supposedly at stake publicly divulged all of the information in the original dataset"* (Millar, 2009, p. 110). Nissenbaum gives an example of the knock-down argument that is frequently used in case law (Nissenbaum, 1998, p. 574). She discusses a case in California Greenwood, in which the Supreme Court ruled that the police did not invade privacy when they asked the Greenwood's trash collector to segregate their trash and turn it over for inspection. This was so, because according to the court majority, people could have no reasonable expectation of privacy to the items that they discarded in an area particularly suited for public inspection (Nissenbaum, 1998, p. 574). Intuitively, this knock-down argument has a lot of appeal. If I make public certain information how can I claim privacy to things that can be inferred from this information? To give another example, if I were to walk around publicly in the city centre, holding hands with a girl my age, it seems unreasonable to expect privacy with respect to the information that the girl is probably close to me. However, in other cases this knock-down argument is not as obvious or effective, which I will discuss in the next section.

2.2. Privacy in public

In the previous section I discussed the contextual account of privacy, and how it allows us to pinpoint some of the privacy concerns raised by data mining. Furthermore, I introduced the knock-down argument, which is commonly used as an objection against claims to privacy with respect to objects or information that are considered "public". Nissenbaum examined the problem of privacy in public extensively, and argues against the knock-down argument, in favour of the possibility and need for privacy in public (Nissenbaum, 1998). Nissenbaum discusses how philosophical theories have imposed limits on the allowable practices of data gathering, analysing and sharing, as attempts to protect privacy, but observed that these limitations are primarily applied to sensitive and intimate information. She argues that there is also a clear relationship between privacy and non-sensitive information that is gathered and analysed from public spheres.

Similarly to Tavani, she identifies two key aspects of public data mining that give rise to privacy issues. The first involves the practice of shifting information from the context in which it was collected to another context. The second involves practices of aggregation, collection and combination of information from various different sources, which could potentially reveal information about an individual (Nissenbaum, 1998, p. 581). Nissenbaum goes into more detail in describing the importance of context with respect to privacy. She discusses the importance of contextual integrity and argues that people more often feel that their privacy is violated by breaches of contextual integrity than with breaches only of sensitive or intimate realms. She illustrates this by arguing that even if information is considered intimate or sensitive, people often do not consider it a violation of privacy to share this information if the information is relevant in the given context. For instance, people usually have no problem with sharing the information about the details of their physical conditions to doctors, or sharing intimate secrets with friends (Nissenbaum, 1998, p. 581). As stated earlier, this contextual nature seems to work well with the access account of privacy.

From this, Nissenbaum draws her main argument, namely that not only information that is considered intimate or sensitive should be protected by philosophical theories of privacy, but also data that is not considered sensitive should be accounted for. She illustrates this by arguing that in practice, information is routinely shifted from one context to another. For instance, when information about an individual's supermarket purchases are sold to a magazine subscription company (Nissenbaum, 1998, p. 585). This example makes sense, as storing information about purchases seems reasonable in the context of a supermarket, but private in the context of magazine subscription companies. This example shows, similar to the example of Tavani, that information that seems not sensitive can still constitute a violation of privacy when accessed in a different

context. In a similar manner she discusses additional concerns with respect to data mining. She argues that data mining practices can be morally questionable, because the data mining process almost always involves shifting information from one context to another context that violates contextual integrity, and secondly, because all the different bits of information combined can reveal private information about individuals quite profoundly (Nissenbaum, 1998, p. 589).

The reader might have noticed that up until now, Nissenbaum's reasoning is similar to Tavani, and the knock-down argument still seems to hold at this point. However, Nissenbaum does attempt to tackle the knock-down argument. She does this by using an example of shopping in a supermarket. Although shoppers in a supermarket implicitly consent to the possibility of fellow shoppers seeing the contents of their shopping carts, they do not implicitly or explicitly agree to other people collecting, sharing and analysing this information (Nissenbaum, 1998, p. 595). Nissenbaum argues that protecting what is valuable in privacy in public involves *"recognizing the distinction between exposing something for observation, on the one hand, and yielding control over it, on the other hand"* (Nissenbaum, 1998, p. 596). To make this distinction between *observation* and yielding *control* more clear another example presented by Nissenbaum is helpful: "you stroll down town wearing a red sweater, then you have freely exposed the information that you were wearing a red sweater at a certain time and date. It is unreasonable to expect that this information may later be suppressed" (Nissenbaum, 1998, p. 572). In other words, it is unreasonable to expect that other people will not look at the clothing that you are wearing in public. However, imagine that instead of just casually observing, someone used the information about you wearing a red sweater to train a machine learning classifier with the aim of uncovering what people of a specific demographic like to wear. This could be considered a violation of privacy, as control has been taken over the information to use it in a different context, which breaks the contextual integrity. The key observation in this example is that no information, even in a public setting, is truly "public", or as Nissenbaum puts it "up for grabs" for data mining (Nissenbaum, 1998, p. 596).

I will now turn back to the main purpose of this chapter, examining how relevant the analyses of Nissenbaum and Tavani are with respect to contemporary data mining practices, specifically, the homosexuality classifier of Wang and Kosinski. Wang and Kosinski state that they obtained the facial images that they used to train their classifier from public profiles posted on a U.S. dating website (Wang & Kosinski, 2018, p. 248). Similar to the distinction between observation and control in the sweater example, the collecting of the facial images from the dating site by Wang and Kosinski breaks contextual integrity. Although the profiles on dating websites are arguably rather "public", anyone who accesses the website can see the profiles, they are made available in a context of observation. When someone creates a profile on a dating website, one does

not implicitly, or explicitly consent to having one's data mined for interesting patterns in dating profiles. Thus, private issues could potentially arise in the training phase of machine learning classifiers, for example in the case of Wang and Kosinski, data from dating profiles was gathered and repurposed for training a homosexuality classifier, breaking contextual integrity.

So, the contextual accounts of privacy of Nissenbaum and Tavani go a long way in pinpointing some of the issues with data mining in relation to privacy. There seems to be a lasting idea of data miners that everything that is publicly accessible on the internet is "up for grabs" and can be used for their data mining practices. However, the contextual accounts of privacy go against the traditional public-private dichotomy of information, by showing that privacy norms are potentially relevant to any information. Data mining practices often break contextual integrity, violating privacy in the process. Lastly, Nissenbaum observes how the aggregation and combination of various pieces of data about an individual can reveal private information about an individual (Nissenbaum, 1998, p. 589).

2.3. Limitations of contextual privacy

Millar states that "even in the wake of their (Tavani and Nissenbaum's) analyses, the knock-down objection remains formidable against their (and any other) contextual account of privacy" (Millar, 2009, p. 110). Although Millar states this, he does not clearly explain why this is the case. In this section I will attempt to do this for Millar. Some pointers as to why Millar believes that the knock-down argument still holds against contextual accounts of privacy can be found in his phrasing of the knock-down argument: *"if individuals have publicly divulged information that is subsequently mined, how can they make a post hoc claim to privacy with respect to information gleaned only from that dataset?"* (Millar, 2009, p. 110).

Millar acknowledges that the analyses of Nissenbaum and Tavani provide important insight into privacy problems associated with the flow of information due to data mining (Millar, 2009, p. 110). As I discussed in the previous section, taking control over information by using it in a different context could break contextual integrity, resulting in a violation of privacy. However, in the discussions of Nissenbaum and Tavani these breaches of contextual integrity are illustrated using examples in which data is used in a distinctively different way than the intended context, such as general data mining for patterns, aggregation of data to create profiles and training machine learning classifiers. But, this shift in context could be avoided by data miners. Taking the homosexuality classifier of Wang and Kosinski as an example, the classifier does not have to be trained on the data of individuals who would consider the use of their data an invasion of privacy. For instance, Wang and Kosinski could have found a number of volunteers, who in the name of

science, voluntarily shared their information with Wang and Kosinski to allow them to train the homosexuality classifier. It is not hard to imagine that data will become such a valuable asset that people are willing to pay money in exchange for information. Solon Barocas and Nissenbaum label this problem the "*tyranny of the minority*: *the volunteered information about the few can unlock the same information about the many"* (Barocas & Nissenbaum, 2014, p. 62). So potentially, patterns can be found, and machine learning classifiers can be trained without breaching contextual integrity, by using the information of people that are willing to share the information.

The question then becomes whether applying the found patterns, or machine learning classifiers to data could constitute a violation of privacy. Or in other words, whether one can make a post hoc claim to privacy to information that is inferred from observing this information. Consider that instead of walking around in a red sweater, an individual decides to post a picture on an online dating profile in which he is wearing a red sweater. Everybody who looked at the dating profile could then conclude that the person wore a red sweater at a certain date and time. However, would it constitute an invasion of privacy if someone was able to infer the price of the sweater, because he worked at a store which sold the exact same sweater? Intuitively, this does not constitute an invasion of privacy because the individual could infer the price of the sweater by merely glancing at the sweater in the picture. However, by slightly changing the example it seems that an invasion of privacy did take place. Imagine that instead of knowing the price of the sweater, an individual used a machine learning classifier which was trained to classify the brand and price of a sweater to make an estimate of the income of the individual. Another example can make this intuitive feeling that privacy can be violated more clear.

Imagine a girl who had grown her hair for a long time with the intention of donating it. Once her hair was long enough it was cut, donated and repurposed into a wig for a girl who lost her hair due to cancer treatments. By doing this, the girl gave access to the information about her hair; what it looks like, what it feels like, the colour, the texture and the smell. Although the girl who received the wig was extremely grateful, out of curiosity she performed a drug test on the hair, from which she found out that the donor was regularly using cocaine. Was the privacy of the girl who donated the hair violated, given that it is plausible to observe drug use based on hair (Boumba, Ziavrou, & Vougiouklakis, 2006), and the girl was actually doing cocaine? Through these examples the limitations of a contextual account of privacy become clear. Although the contextual account of privacy gives us some intuitive understanding as to why this feels like an invasion of privacy, it does not give us a clear understanding. Unlike the contextual breaches that take place when someone repurposes data for mining, it becomes difficult to differentiate between what constitutes a breach of context and what does not. It seems that instead of observing a shift in context, we have

an intuitive feeling that in some cases our privacy has been violated, and hence conclude that contextual integrity must have been breached.

To summarise, in this chapter I examined the works of Nissenbaum and Tavani using Millar's analysis. The analyses of Nissenbaum and Tavani illustrate a major concern with the practice of data mining. Using a contextual account of privacy, Nissenbaum and Tavani show that in the practice of data mining, data is often taken out of context and placed into a different context, violating contextual integrity in the process. This contextual account of privacy holds that information can be public in one context, but deeply private in another. Using this contextual account of privacy Nissenbaum argues against the idea that when information is shared in public places, no post hoc claims to privacy can be made on what is done with this data. She argues that no information is truly out there, "up for grabs", stressing a difference between observing and taking control over data. When we apply this contextual account of privacy to the homosexuality classifier of Wang and Kosinski we find that the process of training their classifier potentially raised privacy issues, because data from dating profiles was used in a vastly different context of training a homosexuality classifier, in which the original data could be considered private. Although the contextual account of privacy does a good job explaining privacy issues that arise due to the flow of data in the process of data mining, it has difficulties articulating the privacy issues in cases when data mining has been done using data that did not invade privacy, and the results of this data mining are applied to new data of individuals to potentially uncover new information about them. Millar suggests that we might be able to articulate more clearly why these inferences, or predictions by data mining feel like an intrusion of privacy, if we focus on the nature of the discovered information, instead of on the privacy issues with respect to the original dataset (Millar, 2009, p .111).

Chapter 3

Core privacy

In the previous chapter Tavani and Nissenbaum's analyses of data mining were discussed. Millar argues that although the analyses of Tavani and Nissenbaum certainly provide important insights into privacy problems that arise during the flow of information in the process of data mining, they are inadequate to provide a full analysis of the unique aspect of data mining, namely the discovery of new knowledge (Millar, 2009, p. 111). Millar attempts to examine whether the information uncovered by data mining can constitute a violation of privacy. Or in other words, whether we can make a post hoc claim to privacy to information inferred from other data. Using the homosexuality classifier as an example, if we are capable of determining someone's sexuality from a picture of his face, could the person still claim that his privacy was violated when he willingly shared a picture of his face on social media and someone used this to determine his sexuality? In this chapter I will give an overview of Millar's analysis of this problem, in which he argues that predictions made by data mining can violate privacy, using his concept of core privacy.

In his analysis, Millar focusses specifically on "complex predictive analyses" in which the goal is uncovering psychological profiles (Millar, 2009, p. 111). He describes how predictive data mining algorithms rely on KDD to extract non-trivial information. In the case of psychological profiling, this includes information about an individual's underlying psychological properties, such as beliefs, desires and intentions. Millar states that it is currently beyond our theoretical landscape to discuss whether predictive data mining could successfully uncover an individual's psychological properties. However, he argues that we can still articulate the privacy implications of data mining practices that aim to uncover our desires, intentions and beliefs. First, Millar starts by discussing the nature of the data uncovered by predictive data mining. He does this using an example in which you have a co-worker named Jon, who shows up at work everyday eating a chocolate doughnut. To illustrate the difference between *descriptive* and *predictive* data mining, Millar states that you could descriptively conclude that "Jon eats a chocolate doughnut every work day", however, predictively you could reasonably draw the conclusion that "Jon likes chocolate doughnuts". Millar discusses how this example illustrates that predictive data mining can uncover information that is qualitatively different from the data in the original dataset.

3.1. The synonymy test

Millar holds that arguments of privacy which claim that the falsity, or the incompleteness of predictions is what makes data mining potentially problematic are easily refuted by only considering predictions by data mining that turn out to be accurate. In order to determine which predictions should be considered accurate, Millar proposes a test which he labels the *"synonymy test"*, which I will briefly outline in this section. Millar argues that the psychological resemblance between the prediction and an individual's underlying psychological properties is most important in an analysis of the success of predictive data mining. If the uncovered data by predictive data mining psychologically resembles an individual's beliefs, intentions and desires, according to Millar, the data mining was successful and the objection that the falsity is what makes data mining potentially problematic can be put to rest. Millar suggest that the psychological resemblance can be assessed empirically.

In his synonymy test, Millar borrows from one of the most well-known examples in artificial intelligence; the Turing test. In a few words, the Turing test suggests that if one is unable to distinguish between machine and human intelligence, this counts towards asserting the machine's intelligence (Millar, 2009, p. 114). Millar proposes that the Turing test can be modified to account for psychological properties. As an example, he discusses a machine that is designed to determine political beliefs. In his example the interrogator is asked to describe his political beliefs in as much detail as possible, this description is then matched against the prediction of the machine. Millar suggests that if the prediction matches the interrogator's own description, then it qualifies as synonymous to his actual political belief. More formally, Millar describes his synonymy test in the following manner: *"if an interrogator is unable to distinguish the emergent data from a self-generated description of the target psychological property of the prediction (e.g. the particular belief, intention, or desire) to a sufficient degree, then the data and psychological property qualify as synonymous"* (Millar, 2009, p. 115).

3.2. Core private information and privacy

Using his synonymy test, Millar examines only the privacy implications of predictive mining that is successful, or accurate. To make this assessment of the privacy implications of emergent data from data mining that is successful, Millar introduces two concepts. The first concept is *core private information*, which he defines as *"an individual's unexpressed psychological properties to which only the individual has first-person access, and that are not knowable by anyone else, except by the individual's prior divulgence of them, or by an unreasonable inference based on other facts already*

known about the individual" (Millar, 2009, p. 117). The second concept is *core privacy*, which he defines as *"A person, P, has core privacy in relation to a piece of core private information, I, and any other person, O, so long as I remains unexpressed by P, such that O can neither observe nor reasonably be expected to infer I"* (Millar, 2009, p. 117). Millar acknowledges that a *"reasonable inference"* is a vague concept to build a theory of privacy on. However, he argues that what can reasonably be expected to infer plays an important role in judging claims to privacy. Millar does propose a definition of reasonable that seems intuitively compelling: *"any inference that an average unassisted person is capable of making given a set of data to which he has access via first-person observation, that is, a person who is not using a data mining algorithm, or working with a trained team of investigators, or referencing a database, etc. is a reasonable inference"* (Millar, 2009, p. 117).

Millar demonstrates his definition of core privacy by extending the example of Jon and his chocolate doughnuts. He concludes from the synonymy test that we would likely consider Jon's desire regarding chocolate doughnuts, and our prediction, that Jon likes chocolate doughnuts synonymous. Millar argues that given the definition of core private information, that the information that Jon likes doughnuts, can not be considered private information with respect to Jon's colleagues, because every average unassisted individual could have made the inference that Jon likes doughnuts based on the information that Jon eats a chocolate doughnut everyday at work. However, he argues that the inference could have been unreasonable if it were made by an employee of Jon's credit card company who could only make the inference because of her access to vast amounts of data collected about Jon and every other customer's purchase. According to Millar, Jon could have a claim to privacy with respect to the information inferred by the employee.

However, this example seems to support the contextual account of Tavani and Nissenbaum more than Millar's concept of core privacy and information. In the example, the privacy problem seems to arise from the fact that the employee uses the information in a manner that violates contextual integrity of the data, because the data was given in the context of making a transaction in order to buy a doughnut, the data may not be accessed with the purpose of determining the preferences of Jon. If we slightly adjust the example we can conclude that there is no privacy invasion considering Millar's definition of core privacy. Consider that Jon is an active user of social media and posts a photo every morning of him eating a chocolate doughnut on the way to work on Instagram. Assume that Jon's Instagram is public, because he loves to share parts of his life with everyone who has access to Instagram. Then everyone who stumbles across Jon's Instagram could make a reasonable, unassisted inference that Jon likes chocolate doughnuts. Looking back at the example given by Millar, the privacy violation arises due to the shift of context from private data

(transactional data), to use in a different context, not because the desire of Jon was core private. The employee could infer that Jon likes chocolate doughnuts by glancing at the original dataset.

Nevertheless, giving a bad example does not take anything away from the usefulness of Millar's analysis with respect to the privacy implications of the unique capabilities of predictive data mining. A better example would be the homosexuality classifier of Wang and Kosinski. Let us stick to Jon, and consider a case in which he has posted a picture of his face, next to a chocolate doughnut that he loves so much. Let us assume that Jon is homosexual. We can furthermore assume that Jon's sexual preference for men could be considered a piece of core private information, as we would generally consider this piece of information a psychological property to which only Jon has first-person access, and that is not knowable by anyone else, except if Jon had made information about his sexual preference public, which in this case he has not. Next, imagine that one of Jon's colleagues used a homosexuality classifier to determine whether Jon is homosexual, from which the result was homosexual. The classifier then predicts that Jon is homosexual, which can be considered synonymous to Jon's sexual orientation according to the synonymity test. Then the use of the homosexuality classifier by the colleague of Jon could be seen as a violation of core privacy, as Jon has not expressed his homosexuality in any way when he was posting a picture of his face on Instagram. The inference by the colleague could not have been made if not for the vast amount of information collected on faces of homosexuals and the machine learning classifier that was trained on this. In no way would we expect that an individual could be able to accurately and reliably predict someone's sexuality. Therefore, Jon could argue that the colleague could not have observed Jon's sexual orientation from his face alone, and he could thus have a reasonable claim to privacy with respect to his homosexuality.

This does, as Millar phrases it "pack an intuitive punch, which goes a long way toward explaining peoples' intuitive objection to data mining" (Millar, 2009, p. 118). I believe that Millar's notion of core private information does not only account for information about an individual's unexpressed psychological properties, despite this already being sufficient to defend a claim to privacy with respect to analyses done by the homosexuality classifier of Wang and Kosinski. We could also make core private information capture a wider variety of information if we consider all data that cannot be observed or reasonably inferred from a piece of data by humans, "core private information". Then the analysis of Millar can cover a broad range of cases such as the girl who donated hair in the previous chapter. It was reasonable of the girl to expect that the receiver of her hair would not be able to reasonably observe or infer the information about her drug use given her hair. Therefore, performing a drug test on the hair constitutes a violation of privacy.

What I find particularly interesting in the analysis of Millar, is that although it seems that Millar acknowledges that essentially core private information can be potentially uncovered by predictive data mining, this does not result in Millar arguing that what should be considered core private information should be shifted. In other words, does it not make more sense to argue that if we are able to accurately classify sexual orientation based on facial images alone, posting a photo of oneself publicly on Instagram would entail giving away the right to claim privacy with respect to one's sexuality? This is were Millar's notion of reasonability comes in. Although it would be difficult to prevent someone from inferring that Jon likes chocolate doughnuts from a photo of Jon with a chocolate doughnut, we could make a moral claim towards people not attempting to infer Jon's sexuality based on a photo on Instagram of his face. Millar's mentioning of the word *"unassisted"* is what makes the difference in here.

To explain this, Nissenbaum's observation on normative claims about why privacy in public is dismissed is useful. As mentioned in the previous chapter, Nissenbaum discusses how it is unreasonable to put restraints on the freedom of others to observe and speak about information that you have made public (Nissenbaum, 1998). However, this does not imply that we cannot put limitations on what can be done with the information. Whereas it would be unreasonable to ask people not to look at the photo of Jon and the doughnut and conclude that he likes doughnuts, it is reasonable to ask people not to analyse your face with the help of a homosexuality classifier, similar to how one can reasonably expect a recipient of donated hair not to do a drug test on it. Perhaps Nissenbaum had already thought about this when she wrote about the distinction between *observing* and *controlling* information. However, Millar's concepts of core private information and privacy help us articulate more clearly what information one can have a reasonable claim to privacy to. Namely, inferred information that goes beyond what the "average" unassisted person could have inferred from observation alone.

There is one critical remark that should be made on the definition of reasonable as proposed by Millar. In his definition he speaks about the "average" person, which is troublesome. For example, in most cases you could reasonably show your friend a picture of you in a sweater, without the friend being able to make a reasonable inference on how much your shirt had costs. However, if I bought the sweater at Primark, and I show a picture of me wearing the sweater to a friend who works at Primark, it becomes more difficult to claim privacy with respect to the cost of the sweater. Millar does briefly acknowledge this problem by stating that the question of reasonableness could be accomplished on a case by case basis (Millar, 2009, p. 118). However, in the case of the homosexuality classifier by Wang and Kosinski, the claim is that the classifier can perceive things in facial images that go beyond what humans could possibly perceive (Wang &

Kosinski, 2018, p. 248). So, in the case of the homosexuality classifier, one's sexual orientation could be considered core private information given only a picture of one's face, regardless of the analytic capabilities of an individual.

To briefly recap, in this chapter I have discussed Millar's concepts of core private information and privacy, to analyse the privacy issues of predictive data mining. Millar goes beyond looking at the context in which the data in data mining is used and looks at the nature of predictions by data mining instead. Millar's analysis helps us in articulating why predictions by data mining intuitively feel like an invasion of privacy; because they have the potential to uncover knowledge that we could have reasonably expected to remain private. In other words, just because it is potentially possible to infer something from a dataset using data mining, does not mean that doing so does not violate privacy. If Wang and Kosinski are right and sexuality is displayed in faces in a way that cannot be perceived by humans but can be perceived by a machine learning classifier, we can make a strong case that uncovering one's sexual orientation using said classifier would constitute an invasion of privacy. Because the information uncovered by the classifier goes beyond what can be observed, or inferred by just glancing at a dataset (in this case a photo of one's face). Nevertheless, the existence of an accurate homosexuality classifier greatly endangers the control that we have over our private information, and our sense of security. However, just because it becomes easier to violate one's privacy, does not take away the fact that one is violating one's privacy.

Chapter 4

The limitations of core privacy

In the previous chapter I discussed Millar's analysis of predictive data mining. Millar's analysis shows us that even though new means to uncover information that would have normally stayed hidden are becoming more available, this does not remove our moral claim to privacy with respect to this information. However, some of the assumptions that Millar makes, severely limit the extent to which his analysis captures the potential problems of predictive data mining. In this chapter I will show that Millar does not analyse the results of predictive data mining in an adequate manner. This, is due to Millar's assumption that the results of data mining can be evaluated in isolation, by focussing on accurate results only. This is problematic because of the following reasons: 1) by ignoring results of predictive data mining that are incorrect, Millar's analysis disregards the potential issues related to false positives and negatives 2) and closely related to this, by examining only cases in which the results are correct, Millar ignores the broader context of predictive data mining. Because of these limitations Millar's analysis is inadequate to answer one of the main questions with respect to predictive data mining, whether the predictive data mining actually uncovers information about an individual.

4.1. Accuracy and the problem of misinformation and privacy

I will start by examining results of predictive data mining that do not pass the synonymy test. Millar's synonymy test does give us some insightful information about the importance of accuracy in predictive data mining. Consider the following example, which was inspired by the famous Harry Potter series[1]. A wizard has brewed a transfiguration potion that allows you to temporarily take on the exact appearance of another person. You take the potion and transform into one of your best friends. Baffled by the power of the potion and your new appearance, and due to the lack of respect for your friend, you decide to run around naked in public, exposing the looks and details of your friends naked body to everybody on the streets. Is the privacy of the friend violated in this scenario?

If the friend considers the information about his naked body private, it seems intuitive to argue that indeed, the friend's privacy is violated in this example. Due to the transfiguration potion of the wizard, accurate information about the friend's naked body is made public, at least if we

1 https://www.imdb.com/title/tt0926084/ (Accessed 7 October 2018)

assume that people on the streets actually looked at your friends naked body. However, there seems to be a turning point at which no meaningful private information is gained. An example that illustrates this are the responses to the emergence of so-called "deepfakes"[2]. Recently, people have been using artificial intelligence, specifically deep neural networks, to take the faces of celebrities and face swap these faces with pornstars in explicit videos. Using this technique people can create convincing videos, in which it really looks like a celebrity is acting in porn. However, although these videos can be very convincing, most of the reactions from the media report the potential security harms that this technology introduces, such as fake news, harassment and hoaxes, rather than seeing the phenomena of deepfakes as a potential privacy threat. It seems that the deepfakes are not considered a violation of privacy because they are not "the real thing". This brings us to the potential harms of false, or misinformation, which Millar ignores in his analysis.

There are multiple problems with deepfakes. One is that they put celebrities in false light, or spread misinformation about a celebrity, by suggesting that they have acted in an explicit video. Secondly, even though it is not actually a celebrity acting in an explicit video, some deepfakes are already convincing enough to make people believe that they now know what a celebrity would look like if he or she starred in an explicit video. Furthermore, even though they nowadays only swap the faces of celebrities and look for pornstars with similar body measurements, in the future the techniques could become more sophisticated, simulating the celebrity even more accurately by using details about the measurements of the celebrity's body, skin colour, tattoos, or whatever information available. However, it would still seem that as long as the actual information about a celebrity's naked body are not available, the explicit video as generated by deepfakes should be considered misinformation, or being put in false light.

The question then becomes whether being put in false light, or having misinformation spread about you can be a violation of privacy. Two recent papers by Pierre le Morvan and Jonathan Schonsheck on this problem are relevant to this discussion (Le Morvan, 2018; Schonsheck, 2018). Le Morvan discusses the debate in information theory on whether information can be false and the implications of this debate for privacy. He identifies two main camps in the debate on whether truth is a necessary condition for information. On one side of the debate people subscribe to information veridicalism which argues that a statement only counts as information if its true, and on the other side you have people who argue for non-veridicalism, which entails that a statement can count as information even if the statement is false (Le Morvan, 2018, p. 81). He then describes different theories of privacy and the effects of siding with either veridicalism or non-veridicalism on ones conceptualisation of the theory of privacy. Since I have started this thesis by assuming that the

2 https://www.kdnuggets.com/2018/03/exploring-deepfakes.html (Accessed 7 October 2018)

access account of information is correct, we only need to look at the consequences of information veridicalism and non-veridicalism on the access account of privacy. If one sides with information veridicalism, being put in false light or having misinformation spread about you would not constitute a violation of privacy. But, if you side with non-veridicalism, being put in false light or having misinformation spread about you could violate your privacy. Le Morvan then discusses how American privacy law compromises four distinct invasions of privacy. One of these is "Publicity which places the plaintiff in a false light in the public eye" (Le Morvan, 2018, p. 84). He then discusses how privacy veridicalism does not cohere with the privacy tort that describes being put in false light, whereas non-veridicalism does.

Björn Lundgren's contributions to the discussion on veridicalism and non-veridicalism can help in the remainder of this thesis (Lundgren, 2017). He argues for a non-veridical conception of semantic information, but at the same time argues that veridical semantic information is still useful as a sub concept of semantic information. The following paragraph is most useful for our discussion: *"We can have an alethically neutral definition of semantic information and, if we want, still claim that the informativity of semantic information depends on truth. Thus, we can deal with the question of informativity as a concept relating to truth without needing to accept any version of the veridicality thesis. The same argument is applicable to contingently false information; i.e. that false information generally is less informative than truthful information does not mean that false information is not information. It just says something about the informativity of false information"* (Lundgren, 2017, p. 13). Thus, rather than questioning whether being put in false light, or having misinformation spread about you is privacy invasive by examining whether the misinformation should be considered information about an individual, it is more fruitful to examine whether the information is *informative* about an individual. Or in other words, if in accessing the information, something private is learned about the individual.

With the previous discussion in mind, I hold, because of similar reasons as Schonsheck (Schonsheck, 2018), that being put in false light, or having misinformation spread about you does not constitute a violation of privacy. Schonsheck discusses that although being put in false light can definitely be harmful, a privacy violation is not a necessary condition for being put in false light. He illustrates this using an example of a male physician of whom false rumours are spread that he sometimes sedates attractive patients and then abuses them sexually. Although this is definitely harmful for the physician, at no point is there any violation of the privacy of the physician (Schonsheck, 2018, p. 99). In other words, the information that the physician supposedly sedates and abuses patients is not informative about the physician in any way. Nevertheless, the example

clearly shows why being put in false light or having misinformation spread about you can still have harmful consequences.

In a similar manner we can evaluate the potential harms of the homosexuality classifier. Imagine that someone is wrongly classified as homosexual, but everyone believes the result of the classifier and holds that he is homosexual. If we take the case of the homosexuality classifier of Wang and Kosinski, the classification was made based on a picture of his face only. For the purpose of this example the wrongly classified individual considers the picture of his face non private information. Then, one could argue that in this case, no violation of privacy has taken place, since the information, the wrong classification, is not informative in any way with respect to the information that the individual considers private. Nevertheless, the misclassification, or misinformation of the homosexuality classifier could have severe harmful implications for the individual. For instance, if a country intolerant to homosexuality were to classify individuals incorrectly as homosexuals, but still acted upon the classification as if they were homosexual, these individuals could face jail, or even be executed without having their privacy violated.

In some situations it might feel like one's privacy is violated by being put in false light. Le Morvan observed this as well (Le Morvan, 2018, p. 86). He discusses how misinformation spread about us can put us into a difficult situation, in which we can only refute the misinformation by revealing private information about ourselves. But he argues, and I agree, that privacy in such a case is only lost when we decide to reveal private information about ourselves, the harm of having misinformation spread itself does not constitute a violation of privacy. Morvan illustrates this using an example in in which Pam publicly accuses Sam, who is sterile, of impregnating her (Le Morvan, 2018, p. 86). In this situation the potential harmful consequences of being wrongfully accused of impregnating Pam can only be refuted by revealing the private information that Sam is sterile.

There is an additional concern in the case of misinformation spread by the homosexuality classifier. For example, consider someone who is put in jail because of being wrongly classified. The individual does consider his sexuality private, but decides to reveal the factual information, that he is heterosexual, to get released from jail. In the best case, the state beliefs the individual and releases the individual at the costs of revealing his sexuality. However, in the worst case, the state could decide to completely ignore the attempt of the individual to reveal his personal information and decide that the classifier is more trustworthy than the individual. This example shows that in both cases, the individual loses control of the ability to share personal information. In the best case the individual is coerced into revealing personal information, whereas in the worst case, attempts of the individual to reveal personal information are not considered at all.

In this section I have discussed the potential harmful effects of false results of predictive data mining. Due to Millar's focus on results that turn out to be correct, the potential harms of false results are not discussed. Admittedly, Millar cannot discuss all the potential harms in the world. However, the discussion on how I hold that having misinformation spread, or being put in false light does not necessarily constitute a violation of privacy, is fundamental to the remainder of this thesis. Although being put in false light can definitely be harmful, this is often the case because of losing control about one's private information, rather than having one's privacy violated. Since I have assumed the access account of privacy as correct, losing control over one's private information does not constitute a violation of privacy. It is therefore essential to examine whether machine learning classifiers such as the homosexuality classifier, actually reveal private information, thus violating privacy; or if they only take away our control over private information.

4.2. The broader context of the machine learning classifier

In the previous section I discussed the potential harms of false classifications. Besides ignoring these harms, Millar's analysis has another flaw due to his focus on accurate results only. Although his suggestion to determine the accurateness of the results of data mining on a case by case basis allows for nuances in what should be considered reasonable with respect to what can be inferred, by examining only single predictions Millar fails to take the broader context into account in which data mining and machine learning takes place. In this section I will show that Millar's approach is inadequate for determining whether a prediction, or classification is a violation of privacy.

The following thought experiment illustrates why this is the case. Imagine that you are homosexual. It is Saturday evening and you are at a large party at the house of one of your best friends. At the party there is someone you are not familiar with and you decide to get to know that person. The person introduces himself as a psychic, claiming to be tremendously good at determining one's sexual orientation. The person stares mysteriously at you for a couple of seconds and then confidently proclaims: "you are homosexual!". You are utterly confused by the claim of the psychic, after all you have done anything you could to hide that you are in fact homosexual. In all your manners, clothing, grooming style and social contacts you attempted to be as stereotypically heterosexual as possible.

Now recall Millar's main assumption for the synonymy test: that if one is unable to distinguish between the prediction and the psychological property to a sufficient degree, the prediction qualifies as synonymous to the psychological property. Then, according to Millar, the data mining was successful and objections that the falsity is what makes predictions by data mining problematic can be put to rest (Millar, 2009, p. 113). If we were to put the individual of the thought

experiment in a room and interrogate him on his sexuality, we would conclude that the prediction by the psychic that the individual is homosexual is synonymous to his own evaluation of being homosexual. Furthermore, the individual did everything he could to hide the fact that he was homosexual. Let's assume that he was successful in hiding this fact and everybody, except for the psychic, always considered him heterosexual. Then in no way has the individual expressed his homosexuality to the public, and hence the individual could reasonably expect privacy with respect to his sexual orientation. Therefore, following Millar's definition of core privacy, this example constitutes a violation of privacy. Continuing the thought experiment, imagine that you then got to know the following:

While you are contemplating how the psychic could have possibly known that you are homosexual your friend screams something to you from the other side of the room: "don't listen to him! The guy is drunk and claims that everybody in this room is homosexual!" Now surely your feelings of having your privacy invaded have faded away. This self proclaimed psychic was just making random claims because of being intoxicated, and coincidently guessed your sexual orientation correctly. Therefore, if we want to make an adequate examination of the privacy implications of data mining, and specifically machine learning classifiers, we should take into account the broader context of the machine learning classifier; how the classifier performs as a whole, rather than evaluating each classification on a case by case basis.

In this chapter I have discussed two problematic aspects of the analysis of Millar. First, Millar ignores the potential harms of inaccurate or false results of predictive data mining. I discussed that in order for uncovered information to be considered privacy invasive, it should be informative with respect to the individual whose privacy is potentially being violated. Second, Millar suggest that we should only evaluate the privacy implications of predictions by data mining that turn out to be accurate, or in in the words of Millar's analysis, that pass the synonymy test. But in doing so, Millar ignores the broader context of the machine learning classifier, the process through which the data mining, or machine learning classifier comes to its prediction. The credibility of a machine learning classifier matters a lot in determining whether something informative is uncovered about an individual. Thus, by focussing only on cases in which predictions turn out to be accurate, Millar ignores one of the main questions at stake, whether predictions by data mining or machine learning classifiers actually uncover informative information about an individual.

Chapter 5

The epistemic value of machine learning classifiers

In the previous chapter I highlighted multiple flaws in the analysis of Millar, most notably that his analysis does not take the credibility of the data mining process into account. Because of this, his analysis is insufficient to adequately examine whether predictions by data mining can violate privacy. Although it seems intuitively correct to consider the accurate uncovering of private information by data mining a violation of privacy, it is only a violation of privacy if the process through which the data mining came to the prediction was credible. In this chapter I will examine the data mining process itself in more detail, focussing specifically on machine learning classifiers such as the one used by Wang and Kosinski. I will show that predictions by machine learning classifiers can potentially violate privacy. In order to do this, I will discuss how we should evaluate the credibility of a machine learning classifier. However, due to the difficulty of interpreting contemporary machine learning classifiers, the danger exists that predictions by machine learning classifiers are not informative with respect to what they are assumed to uncover.

5.1. The nature of machine learning classifier results

In this section I will discuss machine learning classifiers in their broader context, by examining what classifications actually uncover, or access, and how the performance of classifiers is measured. The problem with Millar's approach is that it assumes that machine learning classifications are discrete in nature. Whereas in practice, probabilistic information is generated in multiple stages of the machine learning process.

I will use the homosexuality classifier of Wang and Kosinski as an example, but many of the observations made in this section are applicable to more general machine learning and data mining. Wang and Kosinski discuss the difficulty of interpreting classification accuracy, which they describe as both non trivial and often counter intuitive (Wang & Kosinski, 2018, p. 254). Wang and Kosinski use the area under receiver operating characteristic curve (AUC) coefficient to express the accuracy of their classifier. This AUC represents the likelihood of a classifier being correct when presented with faces of two randomly selected participants, for example an AUC = 0.5 indicates that the classifier is correct half of the time (Wang & Kosinski, 2018, p. 249). Wang and Kosinski warn that

the AUC = 0.91 that they found for their classifier does not suggest that 91% of gay men can be identified, and also not that the classification results are correct 91% of the time.

As in all classification tasks, they discuss how the performance of their classifier depends on a trade-off between what is called precision and recall. In the case of Wang and Kosinski, precision describes the fraction of gay people among those classified as gay, and recall describes the fraction of gay people in the population correctly identified as gay. Wang and Kosinski state that aiming to improve precision reduces recall and vice versa (Wang & Kosinski, 2018, p. 254). They used a logistic regression model, combined with singular value decomposition to train their homosexuality models (Wang & Kosinski, 2018, p. 249). They repeated the procedure of combining logistic regression and singular value decomposition 20 times to assign a probability of being gay to all images in their sample set. Thus, the classifier does not give a discrete answer to the question if someone is homosexual or not based on a photo of their face, but rather assigns a probability, or likelihood of the person in the picture being homosexual. Thus, in order to determine the accuracy of the whole classifier, Wang and Kosinski had to set a threshold for which probabilities are high enough to be assigned the discrete label homosexual. In doing so, they had to make a trade off between recall and precision. They illustrate this using the following example (Wang & Kosinski, 2018, p. 254): Wang and Kosinski simulated a sample of 1000 men and their probabilities of being gay. Since approximately 7% of the population identifies as homosexual, they drew 70 individuals from the set of homosexual participants and 930 from the heterosexual participants. Depending on which probability is taken as threshold, the precision and recall vary. For instance, when Wang and Kosinski took the 70 men with the highest probability of being gay, 39 out of 70 were correctly identified, whereas if they would have taken the 10 men with the highest probability as cut-off point, 9 out of 10 men would be correctly identified as gay.

So, machine learning classifiers do not necessarily generate discrete predictions or classifications. Classifiers often produce probabilistic information instead. To get to discrete results, multiple actions have to be performed, such as determining thresholds and deciding between precision and recall. The question then becomes if the results of machine learning classifiers can still invade privacy if we take into account this process through which the results are constituted. In order to tackle this question I will discuss the practice of machine learning classification in a more general sense, using the concept of profiling, instead of going into more technical details.

5.2. Machine learning classifiers as profilers

Mireille Hildebrandt describes profiling as a set of technologies which are used to create, discover or construct knowledge from huge sets of data (Hildebrandt, 2008, p. 17). She describes how in the

process of profiling, large databases are mined with the aim of finding patterns of correlations between data. Hildebrandt then discusses how profiling is an inductive way to generate new knowledge. Based on correlations or patterns found in datasets, we calculate a probability that things will be the same in the future. As as a working definition she defines profiling as: *"The process of 'discovering' correlations between data in databases that can be used to identify and represent a human or non human subject (individual or group) and / or the application of profiles (sets of correlated data) to individuate and represent a subject or to identify a subject as a member of a group or category"* (Hildebrandt, 2008, p. 19). In other words, a machine learning classifier is trained in order to derive a model, which describes patterns and correlations between features and a desired label. In the case of the homosexuality classifier of Wang and Kosinski, an attempt was made to model how features from facial images correlate to the label of being homosexual.

Hildebrandt distinguishes between two types of profiling: group profiling and personalised profiling (Hildebrandt, 2008, p. 22). She discusses how in personalised profiling, data from diverse sources of one individuated subject is mined with the aim of identifying, and predicting an individual's behaviour. However, in this thesis I am primarily concerned with classifications about an individual that are done on the basis of little information about that individual, such as the homosexuality classifier of Wang and Kosinski, which only requires a picture of ones face to make the prediction, instead of various sources of data. Therefore, I focus on what Hildebrandt describes as group profiling, in which a profile, which has been inductively generated by mining for correlations in a dataset, is applied to a single individual (Hildebrandt, 2008, p. 19).

In this section I have briefly introduced the concept of profiling, the distinction between personal and group profiling and how these relate to machine learning. The concept of profiling is useful as multiple papers have been written about the ethical concerns surrounding profiling using contemporary data mining techniques. In upcoming sections I will use profiles and profiling interchangeably with machine learning models and predictions.

5.3. Machine learning predictions and privacy

With the previous discussion in mind I will now turn to the main question, whether predictions by machine learning classifiers can violate privacy. In this section I will argue that predictions by machine learning classifiers can potentially violate privacy. In the previous section I have discussed how the training of machine learning models is an inductive practice. A problem with models of an inductive nature is that one can never be completely sure if the model will be correct for classifications of future instances. Serge Gutwirth also observed this problem in profiling, and illustrates this using the following example: *"even if the profiling process shows that pattern occurs*

every time some conditions are met, one cannot be 100% sure that it will happen tomorrow as well. Based on its experience, an animal may associate a situation with a danger as a result of recognition of a certain pattern and act consistently, even if the situation in reality, is not a dangerous one: the bad human smell and the shuffling footsteps were not those of a bloodthirsty hunter, but those of a sweet animal rights observer" (Gutwirth & Hildebrandt, 2010, p. 32). Following this line of thought, one could argue that machine learning predictions can never truly violate privacy, because we can never guarantee with 100% certainty that the machine learning predictions will also hold for specific future instances.

An example that does not include machine learning can illustrate the problems with this line of thought. Imagine that you find someone's diary in a busy café. You cannot resist the temptation and decide to read it. While flipping through the diary your attention is drawn towards one quote: "Dear diary, I am not ready to share this with the rest of the world, but I am homosexual". Following the argument against inductive knowledge, I could never be completely sure that the owner of the diary is in fact homosexual. It could be the case that the owner of the diary only wrote down the quote to scare his mother, whom he caught reading his diary, and is in fact heterosexual. Then it seems that almost all indirect assertions, or beliefs that we can hold about individuals on the basis of indirect data could be prone to this line of thought. In order to keep the scope of this thesis clear, I will refrain from going into an in depth discussion about the epistemic validity surrounding beliefs based on inductive knowledge. I will discuss knowledge in the upcoming sections in the everyday use of the word. Meaning that it is reasonable in everyday life to assume that someone who writes that he is homosexual in his diary, is in fact homosexual. As Hildebrandt also observed with respect to this problem: *"How could we move on in life if we did not take certain generalisations for granted, if we did not live by certain rules that are based on such generalisations – even if they do not always apply?"* (Hildebrandt, 2008, p. 24).

Keeping this in mind, I hold, similar to Millar that the predictions of machine learning classifiers could potentially violate privacy. In some cases, even though we cannot guarantee that the results will always be correct, it is still reasonable to hold them as correct (then there is a violation of privacy if the prediction was correct). The discussion in chapter 3 and our current discussion show that their should be an additional criteria for determining whether a prediction is a violation of privacy: both the individual of whom private information is predicted, and the individual who attempts to predict the information should believe that the machine learning classifier could reasonably infer the information. To use the example of the homosexuality classifier, if the classifier correctly determined the sexual orientation of the individual, and both the individual who attempts to determine the sexual orientation and the individual at whom an attempt

is made to uncover his sexual orientation believe that the classifier could have reasonably inferred the information, the privacy of the individual was violated. At least, if the individual had a reasonable expectation that no unassisted human being could have inferred the information (as I discussed in chapter 3). Developers of machine learning classifiers at least claim that they do this, as they attempt to uncover correlations and patterns that are difficult to perceive for human beings, but not for computers. For example the claim of Wang and Kosinski, that machine learning classifiers are significantly better at uncovering sexual orientation than human beings (Wang & Kosinski, 2018).

In this section I have shown that predictions by machine learning classifiers can indeed violate privacy, if the information that was uncovered was considered private and the machine learning classifier is believed to could have reasonably inferred the information, and the inferred information is accurate. However, although one classifier could violate privacy for one individual, it could spread misinformation about another individual. I will discuss the potentially harmful effects of classifiers that do not violate privacy in a later section. Furthermore, as the reader might have observed, there is a limitation in this analysis of how machine learning predictions can violate privacy. In this section I focussed only on discrete predictions that could be verified with a ground truth. In the next section I will discuss whether probabilistic predictions could violate privacy.

5.4. Informative machine learning predictions

As I discussed in the beginning of this chapter, machine learning classifiers often generate probabilistic information at multiple stages in their process, instead of discrete results. In the previous section I discussed how the discrete results could be a violation of privacy. In this section I will examine whether probabilistic information could violate privacy. When machine learning classifiers produce discrete results (or when the probabilistic results are converted into discrete results), it is clear which information they attempt to uncover (or access). However, the probabilistic predictions are often less clear on what they are trying to inform us about.

If a machine learning classifier scans your face, and were to predict that you have a 91% probability of being homosexual, this intuitively feels like newly generated information that violates your privacy. However, this could lead to a belief that I find problematic, that it is possible that all probabilistic information could be seen as potentially privacy invasive. Recall section 4.2 in which I illustrated that having misinformation spread about oneself does not necessarily constitute a violation of privacy, using the example of being wrongly classified as being homosexual. In short, this boiled down to, if you are heterosexual and someone claims that you are homosexual, this does not entail a violation of privacy. Nevertheless this claim could be harmful, due to harmful

consequences or due to losing control over one's private information because of this wrong information. What if instead of someone claiming that you are homosexual, someone states that he believes that you are homosexual with 91% certainty? I would suggest that this statement, is still not a violation of privacy. On the other hand, I just suggested that the probabilistic predictions of machine learning classifiers do feel like a violation of privacy. Therefore, there must be an important distinction between the two cases.

The main difference between the probabilistic prediction by the machine learning classifier and the belief of the individual is the basis on which the claims are made. If someone just states that he believes that you are homosexual without giving any reasons, no private information is accessed, because the belief is not informative about you. In daily life we would except an explanation or arguments that back up the belief. For example, if someone were to believe that you were homosexual on the basis of seeing you in a gay bar, this belief would become a lot more credible. Similarly, machine learning predictions in the case that we are discussing (group profiling), have a pattern, or *model* on which they base their prediction.

Another example that helps clarify the difference is the following. Imagine that you have a small lump in your neck. You read on the internet that a lump could be a potential sign of cancer and decides to visit your doctor. Based on his examination of the lump, the doctor concludes that this lump is likely to be cancerous. Therefore he decides to write an urgent referral for suspected cancer in your medical file. The next day one of your friends goes to the same doctor for a medical check-up. While he is sitting in the doctor's office he notices that the doctor has left your medical file on the bureau. While the doctor is briefly gone to grab some coffee, your friend decides to peek into your medical file reading the referral to the hospital for cancer in the process.

Intuitively this feels like an invasion of privacy. First of all because of the accepted contextual norms that medical information is private information. In addition, your friend has also read something informative about whether you have cancer or not. Even though he does not know the exact observations which has led the doctor to refer you to the hospital, the information that you potentially have cancer is informative in itself. This because we believe in general that doctors are trained and experienced and make reasonable examinations. By changing the example slightly we can clearly see the difference between informative information that can violate your privacy and misinformation. What if you later came to know that the doctor was only pretending to be a doctor all the time, and did not even attended medical school. Then your referral to the hospital is suddenly not so informative and the situation does not feel like a violation of privacy. Similarly, what differentiates a prediction by a machine learning classifier that can invade privacy and one that cannot, is whether the *model* through which they came to there predictions is reasonable or *credible*.

Lastly, imagine that the doctor had valid reasons to believe that you had cancer, but upon further examination it turns out that you do not have cancer. In this case the informativity of the doctor his referral to the hospital because you could potentially have cancer is vastly decreased, illustrating the contextual and temporal nature of informativity.

In this section I have shown that machine learning predictions that are probabilistic in nature can also invade privacy. Furthermore, I have shown the importance difference between misinformation which does not violate privacy, and information that is informative, which could potentially violate privacy. However, both the potential privacy invasions by predictions that are discrete in nature and the ones that are probabilistic in nature depend on the reasonability of the machine learning model in determining whether they can invade privacy or not. Which models should be considered reasonable and which should not will be the question discussed in the upcoming sections.

5.5. Between models and reality

In this section I will discuss criteria for when a machine learning classifier should be considered reasonable, or credible with respect to the generated prediction. An example that illustrates the importance of this question, is the practice of racial science by Nazi teachers in the second world war. Nazi teachers measured external features such as the nose and skill size in an attempt to determine whether students belonged to the true "Aryan race"[3]. How does the homosexuality classifier by Wang and Kosinski differ from these obviously wrong practices? Wang and Kosinski themselves give different examples as to why studying the links between facial features and homosexuality is so controversial (Wang & Kosinski, 2018, p. 246). They discuss how even back in ancient China and Greece decisions were based on facial features, stating that Pythagoras is said to have selected students based on their facial features. Wang and Kosinski discuss how these beliefs have grown in popularity over the centuries, and that even the founder of criminal anthropology, believed that criminals could be identified by their facial features. They place these practices under the label of physiognomy, which they discuss "is now universally, and rightly rejected as a mix of superstition and racism disguised as science" (Wang & Kosinski, 2018, p. 246).

Wang and Kosinski argue that their classifier differs from physiognomy in the following manner: *"Recent progress in artificial intelligence (AI) and computer vision has been largely driven by the widespread adoption of deep neural networks (DNN) … The superior performance of DNNs offers an opportunity to identify links between characteristics and facial features that might be missed or misinterpreted by the human brain"* (Wang & Kosinski, 2018, p. 247). Because the

3 https://www.ushmm.org/outreach/en/article.php?ModuleId=10007679 (Accessed 7 October 2018)

predictions are accurate in a large percentage of the cases, the assumption is that the machine learning classifier has uncovered some meaningful patterns or correlations in the data that were not observable by humans. And this might have some intuitive appeal, as the accuracy with which Wang and Kosinski claim to be able to detect sexual orientation seems daunting at first sight. Furthermore, they claim that the accuracies could potentially be increased by using more data and more sophisticated techniques (Wang & Kosinski, 2018, p. 255). Thus, can the reasonability of a machine learning prediction by judged according to the accuracy of the classifier?

An observation by Hildebrandt helps pinpoint the problem in this approach: *"The caveat of this approach is that it extrapolates from the past to the future on the basis of blind correlations, tending to see the future as determined by established probabilities, possibly disabling potentially better solutions that lie in the realm of low probabilities"* (Hildebrandt, 2008, p. 22). Perhaps the most mentioned concern with respect to big data analytics in general is that correlation does not imply causation (Domingos, 2012; Gotterbarn, 2016; Gutwirth & Hildebrandt, 2010; Macnish, 2017). The problem is that just because correlations can be accurate in describing a relationship between different features and a specific label, this does not necessarily mean that they are therefore reasonable. Macnish describes a number of examples of absurd correlations found by Vigen, for instance that the total revenue generated by arcades correlates with the amount of computer science doctorates awarded in the United States (Macnish, 2017, p. 11).

This discussion shows that accuracy, or strong correlations alone are not sufficient for a model to be considered reasonable. Even in cases where accuracy is close to perfect it is still possible that the correlations behind the high accuracy do not support the claims that are being made in relation to this. For example, consider a large IT company in which the only employed women are cleaners. Even though you could predict with 100% accuracy that a particular woman then is a cleaner, this does not give us any informative information about what it essentially means to be a cleaner. No one would argue based on this information, that women are essentially meant to be cleaners, or are less suitable to hold a job in IT. In other words, this correlation is not informative with respect to these bold statements.

At the same time we must avoid arguing that therefore correlations can *never* be informative information. Although some correlations are spurious, a correlation is often a good indicator of something meaningful, or informative. If placed in an appropriate causal model, classifiers and predictions based on correlations could be informative. Another machine learning classifier that predicts a trait based on facial images, but did not cause the same controversy as the homosexuality, classifier can illustrate this. Zhao discusses how individuals with Down Syndrome have an extra copy of chromosome 21, which is diagnosable by the presence of typical facial appearance and

physical characteristics, such as a small and flattened nose, small ears and mouth and upward slanting eyes (Zhao et al., 2013). Because it is commonly accepted that Down Syndrome can lead to certain external features, it is reasonable to assume that detecting the specific features is informative with respect to whether the individual that is classified has Down Syndrome or not.

As Macnish observed, the problem in the contemporary practices of big data analytics, and specifically machine learning is the tendency to focus exclusively on correlations (Macnish, 2017). Because of this, it is often unclear what machine learning classifiers actually uncover. Furthermore, machine learning classifiers are often so complex, or abstract that it becomes difficult to interpret what they are uncovering. For instance, in order to draw correlations between facial images and homosexuality Wang and Kosinski used a deep neural network called VGG-face, which translates a facial images into 4096 scores (Wang & Kosinski, 2018, p. 249). It is worthwhile to question what the correlations between these features and homosexuality describe. Wang and Kosinski do attempt to go beyond correlations, by suggesting that the correlation found by the machine learning classifier is that faces of homosexual individuals are gender-atypical, which they explain according to the prenatal hormone theory of sexual orientation (Wang & Kosinski, 2018, p. 247). However, as others have observed[4], they do not provide convincing evidence that this is the case. Blaise Aguera y Arcas et al. question whether the machine learning classifier shows that there is a strong correlation between facial structure and sexual orientation, and suggest that it is more likely that the classifier found correlations between more superficial features (Arcas, Todorov, & Mitchell, 2018). To test this, they conducted a survey among 8000 Americans, which asked for their gender, sexual orientation and various other features such as whether someone wears eyeshadow, or has a beard. Based on their survey they found that their was indeed a difference between superficial features among homosexual and heterosexual groups. By creating a simple classification model that asked yes / no questions they could achieve similar accuracies as the classifier of Wang and Kosinski.

So, on closer examination it seems more plausible that Wang and Kosinski have made a classifier that detects the grooming and fashion style in a specific homosexual dating community, instead of finding that homosexuals have a different facial structure due to prenatal hormones. Don Gotterbarn observed exactly the problem that we've seen with the homosexuality classifier of Wang and Kosinski: *"Big data – facilitated a new emphasis on one particular epistemological approach to knowledge acquisition – pattern identification and data analytics – which has led to an unjustified confidence in the truth of claims which are at best conjecture. Because of the quantity and variety of data used in these conjectures they are unjustly elevated to highly probable or even axiomatic level of trust"* (Gotterbarn, 2016). Wang and Kosinski claim to be able to make sweeping

4 http://www.fast.ai/2017/09/13/kosinski/ (Accessed 7 October 2018)

claims about the essence and origin of homosexuality based on faces, while this is not the only explanation of their classifier. However, if it turns out in the future upon further investigation that it can reliably proven that faces of homosexual individuals are in fact significantly different from heterosexual individuals, this could change how we perceive homosexual and simultaneously change our perception on whether it can be predicted. Furthermore, that the machine learning classifier is not informative with respect to you being essentially homosexual or not does not mean that it cannot be informative at all. We should not deny that machine learning classifiers can be trained using amounts of data that would not be processable by human beings. Accordingly, machine learning classifiers can create informative data that could potentially invade privacy. For instance, if Wang and Kosinski's homosexuality classifier indeed uncovered grooming style and fashion, we could still learn informative information about whether someone looks similar to people from a specific homosexual community.

In this section I have discussed the difficulty of determining which claims we can reasonably make on the basis of correlations and patterns found by machine learning classifiers. As many others I have discussed the exclusive focus on correlations in contemporary machine learning practices. Due to the sole focus on correlations, the danger exists of making sweeping claims or predictions that are not necessarily supported by the machine learning classifier. Before we can make claims about violations of privacy, we must first examine what the classifier actually uncovers, or in other words, what the classifier is informative about.

5.6. When models become reality

In the previous section I discussed the difficulty of determining what is actually uncovered in complex machine learning classifiers. Or in other words, what classifications by machine learning classifiers are informative about. I have discussed the risk of claiming that predictions are informative with respect to something, without verifying whether this is actually the case. The problem is that these statements get an elevated epistemic status, even though they do not deserve this status without further investigation. Building on top of the discussions of the previous sections, these claims do not invade privacy in the way that they suggest they do. They claim that they are informative with respect to one's personal information, whereas this is not necessarily the case. However, that does not mean that they are therefore not harmful. Contrary to this, it can be just as, if not even more harmful. In this section I will briefly discuss an example of the potential harmful effects of predictions that are treated as privacy invasions, but are in fact not privacy invasive.

A quote by Gutwirth is relevant for this example: "*This is why we think that profiling is a productive type of knowledge: it tends to create the reality it infers from past occurrences"*

(Gutwirth & Hildebrandt, 2010, p. 32). In order to illustrate this I will use the example of the film 'Minority Report'[5], in which there are mutated humans who are able to see crimes that will happen through visions of the future. Based on these visions, people are arrested before they commit crimes. What if we replace the mutated humans by highly sophisticated machine learning classifiers which predict whether someone will be a criminal or not, should we still arrest the humans who the classifier predicts to be criminals? As the quote of Gutwirth suggests, machine learning classifiers are not passive, rather, they influence and shape the world we live in. If everyone believes that the machine learning classifier can predict whether you are a criminal, this might lead to them treating you differently from others, making you an outcast, ultimately turning you into the criminal the machine learning classifier predicted you to become. Or another less far-fetched example of how machine learning classifiers can influence reality is the homosexuality classifier that we discussed in the previous section. Even if the homosexuality classifier only uncovered whether you dressed like a certain homosexual community, this could still alter the way you dress based on the consequences of this information.

In this chapter I discussed whether predictions by data mining are actually privacy invasive, by examining whether they are informative about an individual. Although it intuitively makes sense to look at isolated predictions, and examine whether they were correct or not, these predictions are only considered informative about an individual when the model through which the data mining, or machine learning came to its prediction is credible. I have shown that accuracy alone is not an adequate measure for determining the credibility of machine learning classifiers. Due to the technical sophistication and the vast amount of data used in data mining, the data mining predictions have acquired an unjust amount of trust, or epistemic value. However, this does not mean that I argue that predictions by data mining could never invade privacy. Quite the opposite, it is hard to deny that through the analysis of vast amounts of data using more and more sophisticated data mining tools new, informative information can potentially be discovered. However, due to the difficulty of interpreting data mining practices such as machine learning, sweeping claims and predictions are made which are not necessarily supported by the machine learning classifier. And these cases in which privacy is not actually violated, can be just as, if not more harmful as when they do violate privacy.

5 https://www.imdb.com/title/tt0181689/ (Accessed 7 October 2018)

Conclusion

I began this thesis by discussing how Wang and Kosinski believe that due to the growing digitalisation of our lives and the rapid progress in artificial intelligence, we are inevitably headed towards a world in which privacy has been completely eroded, which they label the *post-privacy* world. At first sight, the results of their homosexuality classifier make a compelling case for believing that intimate traits can be uncovered using artificial intelligence, potentially endangering our privacy. In order to examine this post-privacy narrative, this thesis examined the following question: "*Can predictions by machine learning classifiers violate one's privacy?"*.

To answer this question, I first examined contextual accounts of the privacy implications of data mining. The contextual accounts go against the traditional public-private dichotomy of information, by showing that privacy norms are potentially relevant to any information. Using the analyses of Tavani and Nissenbaum I discussed one major concern with data mining. In order to find patterns in large amount of datasets, or to train sophisticated machine learning classifiers, data is often shifted from the context in which it is shared, to a context of analysis. However, data could be public in one context, but deeply private in another. I applied this to the homosexuality classifier of Wang and Kosinski, and argued that the individuals whose data was used to train the homosexuality classifier could claim that their privacy was violated. The individuals willingly shared their sexual orientation and a photo of themselves on a dating website to find potential dating partners, but could still find it privacy invasive if this data was used to train a homosexuality classifier.

Next, I discussed how the contextual accounts of privacy have difficulties with articulating limitations on what can be done with data. Although it is compelling that repurposing data for training a homosexuality classifier constitutes a violation of privacy, it is less clear in other cases. If it is possible to infer private information from public information using machine learning, is it still reasonable to expect privacy with respect to the inferred information, if the public info was willingly shared and the machine learning classifier was trained in a manner that did not violate privacy? This problem lies at the heart of the assumption of Wang and Kosinski that we are inevitably headed towards a world in which privacy is completely eroded. It suggests that if machine learning is in fact able to infer private information from public information, then the predictions by machine learning would not violate privacy in these cases, rather they would shift what is considered private information. However, using Millar's concepts of core private information and privacy I argued that we could have a reasonable claim to privacy with respect to

predictions by machine learning classifiers. Namely, in cases in which the predicted information goes beyond what the average unassisted person could have inferred from observation alone. Nissenbaum discusses something similar to this in her discussion about *observing* and *controlling.* However, by focussing on the nature of the predictions rather than the public-private status of the original dataset, Millar articulates more clearly why these predictions could be privacy invasive. Thus, Millar shows us that even if it is possible to predict sexuality from facial images in the way Wang and Kosinski suggest, it could still be considered a violation of privacy if the classifier was used to predict the sexuality of someone who publicly shared a facial image.

Despite its usefulness, I discussed some limitations of the analysis of Millar. Due to Millar's sole focus on accurate results, he fails to take into account the broader context of the data mining process. Predictions by machine learning classifiers cannot be evaluated in isolation, the overall credibility of a machine learning classifier is essential in determining whether something informative is uncovered about an individual. I discussed how I hold that misinformation does not constitute a violation of privacy. Although it is possible that control over private information is diminished when misinformation is spread about an individual, the privacy of the individual is only violated when private information about the individual is accessed. I then turned to the main question whether predictions by machine learning classifiers can actually uncover private information about an individual. Specifically, whether the homosexuality classifier of Wang and Kosinski truly observed informative information about one's sexual orientation from facial images, that is not perceivable by humans. In the paper of Wang and Kosinski, no convincing evidence could be found for the claim that the machine learning classifier has uncovered a pattern about sexual orientation in facial images that is not perceivable by humans. It is more likely that the classifier learned to identify how a specific demographic grooms itself than that it has found a deterministic model between facial features and sexual orientation. Nevertheless, we cannot exclude the possibility that, through the analyses of vast amounts of data, machine learning classifiers could potentially make informative predictions that cannot be made by humans.

The problem that we face nowadays in contemporary machine learning practices, such as the homosexuality classifier of Wang and Kosinski, is the difficulty of interpreting what the predictions are informative about, or in other words, what the machine learning classifier has learned. However, due to the technological sophistication and the vast amount of data used in machine learning, the predictions have acquired an unjust amount of epistemic status. The danger exists that sweeping claims (such as that the homosexuality classifier of Wang and Kosinski is able to detect hidden patterns in facial information that determines sexual orientation) are assumed to be true, even when there is no convincing evidence that this is the case.

Now to give a concluding answer to the main research question: *Yes, predictions by machine learning classifiers could potentially violate one's privacy.* First of all, because in order to make the predictions machine learning classifiers have to be trained, which is often done using data that is taken out of its context, breach contextual integrity and privacy in the process. Furthermore, the existence of a machine learning classifier that could uncover private information does not take away the reasonability of a claim to privacy with respect to this information. Last, although privacy could potentially be violated by predictions of machine learning classifiers, it could be just as harmful, perhaps even more harmful, when predictions do not violate privacy.

To end this thesis, I will discuss some of the limitations of this thesis and suggest directions for future work. First a practical limitation. In chapter 3, I discussed that even if it is possible that private information is uncovered about an individual using a machine learning classifier and a piece of information that was willingly shared, an individual can still claim that their privacy has been invaded if it is actually uncovered. However, would this matter in practice? I agree with Wang and Kosinski that the internet is difficult to police. Will our claim to privacy still remain when more and more control about our information is taken away?

One of the main contributions of this thesis to the debate between Tavani, Nissenbaum and Millar is, that the broader context of the data mining process should be taken into account. In chapters 4-5 I argued that predictions by machine learning classifiers can only violate privacy, if the model, or the machine learning classifier through which the prediction was done is considered credible. However my analysis of what should be considered credible is quite shallow. Rather than given a clear account of what should be considered credible, this thesis illustrated the importance of taking into account the credibility, by pinpointing cases which are clearly not credible. Future research could delve further into the problem of interpreting machine learning classifiers, and delve into the epistemological debate on what should be considered a credible machine learning classifier with respect to what it aims to uncover.

Last, the ethical outworking of the observed problems in this thesis are a good starting point for future work. Although I briefly outlined why predictions of machine learning classifiers that do not violate privacy, but hold the same epistemic status as those who do can be harmful in section 5.6, there is definitely more to say. Wang and Kosinski seem to believe that machine learning is a tool which can be used to passively observe the world, to uncover the underlying workings and patterns of reality, whereas in practice, predictions by machine learning classifiers actively help in determining what reality becomes.

References

Arcas, B., Todorov, A., & Mitchell, M. (2018). *Do algorithms reveal sexual orientation or just expose our stereotypes?* Augmenting Humanity. Retrieved from https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around anonymity and consent. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, *1*, 44–75.

Boumba, V. A., Ziavrou, K. S., & Vougiouklakis, T. (2006). Hair as a biological indicator of drug use, drug abuse or chronic exposure to environmental toxicants. *International Journal of Toxicology*, *25*(3), 143–163. https://doi.org/10.1080/10915810600683028

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

Gotterbarn, D. (2016). The creation of facts in the cloud: a fiction in the making. *ACM SIGCAS Computers and Society*, *45*(3), 60–67.

Gutwirth, S., & Hildebrandt, M. (2010). Some caveats on profiling. In *Data protection in a profiled world* (pp. 31–41). Springer.

Hildebrandt, M. (2008). Defining profiling: a new type of knowledge? In *Profiling the European citizen* (pp. 17–45). Springer.

Jernigan, C., & Mistree, B. F. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, *14*(10).

Le Morvan, P. (2018). Information, Privacy, and False Light. In *Core Concepts and Contemporary Issues in Privacy* (pp. 79–90). Springer, Cham. https://doi.org/10.1007/978-3-319-74639-5_6

Lundgren, B. (2017). Does semantic information need to be truthful? *Synthese*, 1–22. https://doi.org/10.1007/s11229-017-1587-5

Macnish, K. (2016). Government Surveillance and Why Defining Privacy Matters in a Post-Snowden World. *Journal of Applied Philosophy*, *35*(2), 417–432. https://doi.org/10.1111/japp.12219

Macnish, K. (2017). Taking Shortcuts: Correlation not Causation, and the Moral Problems it Brings. Retrieved from https://www.academia.edu/34694031/Taking_Shortcuts_Correlation_not_Causation_and_the_Moral_Problems_it_Brings

Millar, J. (2009). Core privacy: a problem for predictive data mining. *Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society*, 103–119.

Nissenbaum, H. (1998). Protecting privacy in an information age: The problem of privacy in public. *Law and Philosophy*, *17*(5–6), 559–596.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 67.

Schonsheck, J. (2018). The Unrelenting Darkness of False Light: A Sui Generis Tort. In *Core Concepts and Contemporary Issues in Privacy* (pp. 91–106). Springer, Cham. https://doi.org/10.1007/978-3-319-74639-5_7

Tavani, H. T. (1999a). Informational privacy, data mining, and the Internet. *Ethics and Information Technology*, *1*(2), 137–145. https://doi.org/10.1023/A:1010063528863

Tavani, H. T. (1999b). KDD, data mining, and the challenge for normative privacy. *Ethics and Information Technology*, *1*(4), 265–273. https://doi.org/10.1023/A:1010051717305

Venkatadri, M., & Reddy, L. C. (2011). A Review on Data mining from Past to the Future. *International Journal of Computer Applications*, *15*(7), 19–22. https://doi.org/10.5120/1961-2623

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. https://doi.org/10.1037/pspa0000098

Zhao, Q., Rosenbaum, K., Sze, R., Zand, D., Summar, M., & Linguraru, M. G. (2013). Down Syndrome Detection from Facial Photographs using Machine Learning Techniques. In *Medical Imaging 2013: Computer-Aided Diagnosis* (Vol. 8670, pp. 867003-1-7). International Society for Optics and Photonics.