The Disciplinary Power of Algorithms: Domination, Agency and Resistance

by

Kelly Mink

A Master thesis in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

MA Applied Ethics
Faculty of Humanities
Utrecht University
June 2020

Student number: 6853617
Supervisor: dr. Sven Nyholm
Second reader: Siba Harb

Abstract

In our daily lives, we are regularly unaware that we are confronted with algorithmic decision-making systems that use Big Data and machine learning to improve their efficiency and accuracy. Algorithms do not merely increase efficiency; they can also be used to mediate social processes, construct your identity, and can create the opportunity for normalization. However, they can perform these tasks without adequate transparency and accountability. The subjection of the individual to such systems raises the question of whether we are in some way dominated by those systems. Therefore, this thesis aims to answer two critical questions: whether we are dominated by algorithmic decision-making systems, and if we are, what resistance against this domination should look like. Using Foucault, a neo-republican account of freedom as non-domination as used in surveillance studies, and the concept of 'micro-domination', I argue that we are indeed dominated by those automated systems, but should extend the scope to the collaborative relationship between the system and the human agents involved. Resistance against this micro-domination should at least (1) uncover the asymmetrical power relations involved in the decision-making process, (2) identify the relevant agents involved, (3) incorporate democratic values and track citizen's interests to empower them and (4) make the overall system more transparent to the individual. Furthermore, this thesis tests whether the theory of meaningful human control over automated driving systems could satisfy these conditions. However, it shows that it is unable to fully overcome the problem of distribution of responsibility amongst relevant human agents. I introduce the term 'micro-resistance' within the context of algorithmic decision-making, which, despite its smaller scale, may have a significant impact on the criticized system when structurally imposed.

Table of Contents

Introduction

Jobhunting today is a different practice than it was a decade ago. In this new era of digital technologies, career sites like LinkedIn are increasingly popular, whose algorithms consistently select vacancies that suit your interests based on your profile, online 'likes' and scrolling behavior. Furthermore, companies seeking new employees make more and more use of algorithmic decision-making systems to screen vacancies' applicants and determine who is suitable for the job and who is not (O'Neil 2017). The use of algorithms in these instances makes it seem like I have to make sure that not the human beings behind the screen, but the algorithm used regards me as the best candidate for the job.

What does it mean that the possible decision to invite me to a job interview seems to lay not in the hand of a human being, but of an algorithm? What are the implications of this technological transition? We are confronted with decision-making algorithms more often than we might think, and the decisions they are making are becoming more influential in an individual's life. They are finding their way not only in job applications but also in online advertising, applying for a loan, and even in our criminal justice system (O'Neil 2017; Martin 2019). Algorithms mediate social processes, governmental decisions, and how we see ourselves and our environment (Mittelstadt et al. 2016). Furthermore, their decision-making strategy is often opaque to those who are affected by it.

The impact of algorithms on our lives may be higher than we realize. This impact requires us to face the issues from a multidisciplinary standpoint, including law, computer science, and the ethics of algorithms (D'Agostino and Durante 2018). The combination of accountability questions and their growing influence on our day to day lives also give rise to a more politically oriented question: are we being dominated by those algorithmic systems?

The research question that this thesis aims to answer is twofold: are we, individuals who are subjected to opaque decision-making algorithms, being dominated by those algorithms, and if we are, what should resistance against this domination look like?

In this thesis, I will argue that we are micro-dominated by algorithmic decision-making systems and that resistance should take the form of an anti-power that empowers the individual and ensures that the system becomes more democratized, transparent and easier to understand by those affected by it. I will call this form of resistance 'micro-resistance'.

As an introduction to the subject, chapter one will focus on algorithms in general. It will answer the question of what algorithms are and what problems algorithmic decision-making face. I will argue that a significant factor that makes algorithmic decision-making a form of domination is the accountability issues. Therefore, chapter two will focus on accountability issues in algorithms, focusing on opacity and possible obfuscation of responsibility by relevant human agents. Chapter three will focus on domination. In this chapter, I will analyze the concept of domination using Foucault and place his concept in a neo-republican framework that regards freedom as non-domination using literature from surveillance studies. I will introduce the term 'micro-domination' as put forward by Danaher (2019) and O'Shea (2018) and critically assess whether it is possible to ascribe domination to a non-human agent using Nyholm's concept of collaborative agency.

Chapter four will be dedicated to the subject of resistance, answering the second question of this thesis. I will argue that resistance should take the form of an anti-power and set four preconditions for resistance focusing on uncovering the power relations, identifying relevant agents, and incorporating democratic values and transparency. I will test the theory of meaningful human control over automated driving systems as put forward by van den Hoven and Santoni de Sio (2018), extended by the latter and Mecacci (2019), as a possible form of resistance when applied to the context of algorithmic micro-domination. While the theory proves to be insufficient, I will use relevant components of the theory to argue for what I will call micro-resistance.

I. Algorithms

1.1 What algorithms?

When attempting to discuss an 'ethics of algorithms', is it necessary to explain what algorithms are and specify what kinds of algorithms we are going to discuss. There are many possible explanations of the term 'algorithm', varying from the technological explanation to a more social-technological interpretation. According to Kraemer et al., an algorithm is a '*finite sequence of well-defined instructions that describe in sufficiently great detail how to solve a problem*' (2011). Another explanation, more fundamentally, is that an algorithm is '*nothing more than a very precisely specified series of instructions for performing some concrete task*' (Kearns and Roth 2020, 4).

It is important to stress that my focus will be on algorithms that influence your life on a relatively small scale. Most of the literature focusing on the problem of domination and responsibility in machine-learning algorithms focus on systems that have to decide between life and death, like in autonomous driving systems, or the idea of 'Killer Robots'. Whether or not those autonomous systems are dominating us tends to sound like science-fiction story plots, but this is not the scale of systems that I would like to address. Therefore, the overall aim of my thesis is to investigate whether or not we are being dominated by algorithms that we are confronted with right now in our daily lives, especially stressing the importance of critically assessing their role and the influence they have on us and the lack of control we have over them.

I distinguish two types of algorithms that desire individual attention. The first type of algorithm is what I will call 'simple' algorithms. It is an algorithm that is fundamental in that it does not show any sign of intelligence or autonomy. These algorithms merely perform the tasks exactly as put forward by their designers, such as sorting a list of contestants in alphabetical order. The second type of algorithm is what I will call 'smart' algorithms. They are algorithms that use machine learning and artificial intelligence to learn independently from the algorithm's designer. While a human being is involved in designing the algorithm, she does not directly design the final algorithm herself. The

final algorithm is derived from data and meta-algorithms, and its output is itself another algorithm that can be applied to further data (Kearns and Roth 2020, 6).

1.2 Why use algorithms?

Simple algorithms are used primarily to increase efficiency. If we imagine the task of sorting a list of contestants in alphabetical order, it is much easier and less time consuming to design an algorithm that can do the task for you instead of doing it yourself, especially if the size of the data set increases over time. Smart algorithms, on the other hand, are used to get algorithms to perform more complex tasks. According to Kearns and Roth, they ensure that we can use functions like face recognition, language translation and advanced predictions primarily based on data and machine learning experience (2020, 6). Smart algorithms can turn large data sets into newer, even smarter algorithms that can detect patterns in data that cannot be detected by human beings.

The use of these large sets of data, also referred to as Big Data, is especially interesting because they can identify patterns and correlations, converting massive volumes of data into a particular, highly data-intensive form of knowledge (Cohen 2012, 1919). Companies can use these highly advanced algorithms to improve and simplify their work. For example, advertisement companies can use these smart algorithms to find patterns in an individual's behavior to adjust their advertisements to reflect their costumers' interests and wishes as much as possible, increasing the business's profit. Smart algorithms ensure that we have more tailored news, better traffic predictions, more accurate weather forecasts and better suiting search results (Martin 2019, 836). They can even have predictive capacities. A famous anecdote to show how advanced these algorithms can get is how Target predicted a girl was pregnant based on her buying behavior, before telling it to her parents (Hill 2012). Although this 'proof' is anecdotal, it displays the possibilities of these machine-learning algorithms, and what kind of information these algorithms can give to companies who are interested in them.

Sometimes, there is more to algorithms than merely a tool to optimize profit or make decisions. Some argue that algorithms can also be used to regulate behavior. According to Karen Yeung, there is a distinction between algorithmic decision-making and algorithmic regulation (2018, 3). Algorithmic decision-making entails that a decision is being made aided by algorithmically generated knowledge, whereas algorithmic regulation focuses on regulatory systems that utilize algorithmic decision-making. This could indicate that algorithmic systems can also be used as a form of regulation. She refers to algorithmic regulation as '*decision-making systems that regulate a domain of activity in order to manage risk or alter behavior through continual computational generation of knowledge from data emitted and directly collected from numerous dynamic components pertaining to the regulated environment in order to identify and, if necessary, automatically refine the system's operations to attain a prespecified goal'* (2018, 507*).* John Danaher uses Yeung's definition of algorithmic regulations in his conceptualization of an 'algorithmic tool', arguing that such a tool is used to 'generate and act upon knowledge that can be used to aid decision-making and goal achievement, and that functions in part by regulating and governing behavior' (2019, 101).

Jack Balkin also claims that algorithms do much more than merely making processes more efficient. He argues that "*algorithms (a) construct identity and reputation through (b) classification and risk assessment, creating the opportunity for (c) discrimination, normalization, and manipulation, without (d) adequate transparency, accountability, monitoring, or due process"* (2018, 1239). This makes the use of algorithms in decision-making systems also very interesting from a philosophical and ethical perspective, as this thesis will demonstrate.

1.3 The problem of bias and accountability

One major possible advantage of algorithmic decision-making over human decision-making is the removal of bias in decision-making, claiming that algorithms are 'neutral' (Noble 2018, 1). A company could use such algorithms to let technology choose who gets invited to a job interview instead of letting a human being decide to 'ensure' that it is not influenced by human emotions. However, the use of algorithms to make decisions

does not necessarily remove bias from the process. Using algorithms to decide who gets invited to the job interview does not mean that the output of the algorithms will never be biased in a sense. An 'objective' and 'unbiased' algorithm can still produce biased outcomes if the input is still biased. Using an algorithm with the supposition of removing bias from the decision is not useful if you do not explicitly "remove" the bias from the algorithm, which includes explicit and implicit bias.

Although the problem of biased algorithms is not the main subject of this thesis, it does pose interesting questions regarding responsibility issues. I cannot see the sets of rules that were assigned to the algorithm that makes a decision, and I also have no direct insight into the workings of the system overall. Therefore, it is the question whom I can hold accountable in case I have the feeling that the algorithm may be biased in a way that negatively affects me. Imagine the example of the job application, and I get turned down by the company but also find out that all the other women who applied also got turned down. I may wonder if the decision was based on a criterium like gender instead of capacities. This could either be an explicit bias (a direct command from the hirer of the company) or an implicit bias used by the algorithm to judge me (all past successful employees in this position were men). Who can I hold accountable for the workings of the algorithm? The next chapter will focus on this problem in algorithmic decision-making precisely.

II. Algorithms and Accountability

If I do not agree with the decision-making procedure during my job application, the most obvious thing to do is ask the company why I got rejected. In the case of human decision-making, the company will probably connect me to the person in charge of human resources and application procedures. That human will give me his or her reasons, aligned with the company's policies and interests, to justify my rejection. In the case of algorithmic decision-making, this procedure may be impeded. Who is responsible for the decisions made by algorithms?

This short chapter will focus on accountability in algorithmic decision-making. Apart from it being a genuine problem within algorithmic decision-making systems and a matter that continuously pops up in AI literature, some problems contribute to the feeling of domination by those algorithmic systems. Therefore, this chapter will discuss accountability issues in the amount of detail that I could or that it, perhaps, deserves[1]. However, it will present pressing issues with accountability that will substantiate the claim of algorithmic domination and provide specific issues that a possible form of resistance should incorporate.

Within philosophical traditions, there are various interpretations of what responsibility and accountability entail, and since this chapter is not the core argument of this thesis, I will not go into much detail regarding the various interpretations possible. However, it is necessary to say something about the difference between (moral) responsibility and accountability. I will use Rubel, Castro, and Pham's structure of responsibility for my understanding of the concept when necessary. It claims that moral responsibility is a conjunction of role responsibility, causal responsibility, and capacity responsibility

[1] Some references to a more extensive discussion of accountability include Fischer and Ravizza (1998; 2012) on moral responsibility and reasons-responsiveness, Martin (2019) on accountability in algorithms, and Binns (2018) on accountability in machine learning.

(2019, 1020)[2]. My focus is on accountability, with which I aim to not only identify who is morally responsible but also what that responsibility involves (Rubel, Castro, and Pham 2019, 1021). I see the distinction between the two as follows: in my interpretation, moral responsibility involves what the agent ought to do, what her moral obligations are in a particular context, and requires the moral responsibility framework as written above. Accountability describes what that responsibility involves - in my view, accountability is the ability and the willingness (or, perhaps, the obligation) to justify one's actions or the actions relevant to that agent in virtue of their role. So, applied to the example of the company who has to decide who to hire, the company's moral responsibility may involve letting the procedure be fair to all the applicants, and to treat everyone with respect and dignity. Holding the company accountable for the outcomes of the procedure means that there is someone (or multiple agents) who are willing to give me reasons and justifications for the procedure. So, while the two concepts are closely intertwined, they have different focal points and raise different issues. Furthermore, my concept of responsibility is compatibilist to the extent that I do not believe that a delegation of decision-making to non-human agents amounts *per se* to the disappearance of human moral responsibility for these decisions (Santoni de Sio and van den Hoven 2018, 5)[3].

2.1 The Black Box society and responsibility gaps

As stated earlier, machine-learning algorithms behave differently from 'simple' algorithms – they can learn independently from the programmers that created them. In that process, they produce newer, smarter and more specified algorithms to improve the

[2] Rubel, Castro and Pham use Hart (1968), Vincent (2011) and Kutz (2004) for their account of the structure of responsibility. Role responsibility refers to specific responsibilities that occur in the virtue of the role of the agent, in which certain events within the domain should be anticipated and taken action for when necessary. Hart uses the example of a captain who is responsible for the ship because she is the captain (1968,211). Causal responsibility is the link between the agent's action and the event that results from it. Capacity responsibility relates to whether an agent has the requisite capacities to be responsible for an outcome, related to whether the agent has enough information and is not suffering from non-deliberative action for which she lacks the required capacity to be responsible (Rubel et al., 1019-1020).

[3] Santoni de Sio and van den Hoven use Fischer and Ravizza (1998)'s view on accountability, arguing that moral responsibility requires an agent to exercise *guidance control.* A more extensive explanation of their view and the content of guidance control will be presented in chapter four on meaningful human control.

outcome of their prescribed task. However, the process of creating these new algorithms is opaque. This process is often referred to as a 'black box', meaning that we only know the input of the algorithm and the outcome of the process while the process itself remains hidden from the public eye to see. Frank Pasquale uses the term 'black box' in an algorithmic society for its dual meaning: it can refer to a recording and data-monitoring device (like used in airplanes), and a system whose workings are mysterious (2015, 3). It seems to make sense that not knowing what is inside the black box is closely linked to the problem of attributing accountability. However, apart from the question of whether we can open the box (technologically), it is also interesting to investigate whether companies would actually *want* to open the box.

Pasquale argues that there are three strategies for keeping black boxes closed: 'real' secrecy, legal secrecy, and obfuscation (2015, 6). Real secrecy involves a barrier between content and unauthorized access, like using a password to protect unauthorized people from entering your laptop. Legal secrecy is the obligation to keep certain information secret, for example; the obligation my doctor has to keep my medical information private from external companies. And obfuscation, the most interesting one to me, is the deliberate attempt at concealment when secrecy has been compromised. These three together form his term for *'remediable incomprehensibility'*: opacity (2015, 6).

Obfuscation is a method that we all have encountered often in our digital lives: I believe that we can find one of the most profound examples of obfuscation when accepting the terms and conditions of a specific service or product. If we want to take a sneak peek in some parts of the black box of a specific service, some details can be found in the terms and conditions available to you, and to which you have to agree before you can start using the service. This can be regarded as a case of obfuscation because the terms and conditions are extremely long and often difficult to really understand without a specific background in law, data science or any other related field. Therefore, companies can argue that they are already transparent about their data usage and intends with the product by presenting the terms and conditions. However, by deliberately making the process difficult for you to understand, it can be interpreted as an attempt to prevent you from actually reading and understanding what is going on. Rubel et al. describe a

similar phenomenon as 'agency laundering', in which agents obscure their moral responsibility for the technological systems' results inside their decision-making processes (2019, 1018). When an agent launders her agency, she uses the automated process as a disguise to distance herself from morally suspect actions that she actively attributes to the automated system, which can be seen as a moral wrong (2019, 1021).

What could be reasons for companies not to be willing to open their black boxes? Part of this has to do with business secrecy and the business model of algorithms. If you have a very successful algorithm, it makes sense that you are not willing to give up your secret weapon to the public. Sometimes, secrecy may even be warranted, for example in the context of preventing terrorist attacks and protecting national security (Pasquale 2015, 4). In most cases, the only way in which we actually find out what is going on is in the case of whistleblowers like Edward Snowden. He reported on the NSA's secret surveillance practices in 2013, which was the biggest intelligence leak in the NSA's history (Snowden 2019).

Apart from instances in which agents deliberately try to obfuscate their responsibility, there are also cases – especially in machine-learning algorithms - in which genuine responsibility gaps seem to occur. One context in which responsibility gaps can occur, and which is the most relevant in our case of algorithms, is when nobody can reasonably be held responsible or accountable due to the opaqueness of the system, making its workings and its consequences unpredictable (Santoni de Sio 2016, 20). Within algorithmic systems, the allocation of responsibility is especially interesting if we attribute a specific value to the algorithms involved in the system, making them value-laden. Some authors argue that at least some specific algorithms are value-laden, meaning that the developers of the algorithm incorporated some of their ethical values into the algorithmic systems (Kraemer, van Overveld, and Peterson 2011; Mittelstadt et al. 2016; Martin 2019). When delegating the decision-making process to an algorithmic system, multiple questions arise: can we assign a certain amount of responsibility to the system, which is a non-human agent? Can we even attribute agency to the system in such a way? And if we cannot, who can be held accountable for the workings of the system?

2.2 The ideal of transparency

An idea that can initially be proposed when dealing with responsibility problems in algorithms is the ideal of transparency: if we can turn the opaque system into a transparent system, we know more about the decision-making process, which can help us identify the parts of the system that influenced the decision and find the relevant agents involved in the design of those parts of the system. In short: idealistically, transparency gives me the tools to hold someone accountable for the workings of the system. Pasquale proposed that we may need to ensure that algorithmic agents are traceable to their creators directly (2015, 1253). Will transparency be the solution to our problems?

Burrell is critical of the ideal of transparency and distinguishes three forms of opacity:

1. Opacity as intentional corporate or institutional self-protection and concealment.
2. Opacity stemming from the current state of affairs where writing and reading code is a specialist skill
3. Opacity that stems from a mismatch between how computers and mathematics work in machine learning and the way human beings reason and use semantics to interpret things (2016, 2).

The first form of opacity is intentional and has been covered to some extent above. The third form of opacity is concerned with the complicity of machine learning in general, especially when it has to handle a massive amount of data. Burrell argues that *'while datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity (and thus opacity)'* (2016, 5). Last, the second form of opacity is why I believe only making opaque systems transparent is not enough, especially not in our example of the system that determined that I was not eligible for the job I applied for. I would need to be able to interpret the algorithmic system and be able to have some understanding of what is going on to make the insight valuable to me in any sense.

Ananny and Crawford agree to this, writing that we should hold systems accountable not by looking at what is on the inside, but by looking across them – seeing them as '*sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans'* (2018, 2). Furthermore, they are critical of transparency as an ideal, presenting several negative implications like creating false binaries and do possible harms. So, while transparency may be useful to display asymmetrical power relations (of which the importance will be discussed in the upcoming chapters), it is not enough to solve the problem overall due to responsibility problems, epistemic shortcomings, and the limits of the transparency ideal. Therefore, we can regard transparency as a step in the right direction, but not necessarily one that satisfies our desires to the full extent.

To summarize, it looks like our daily lives are increasingly influenced by self-learning algorithms that suffer from accountability issues and lack a satisfying form of transparency. This can result in a feeling of powerlessness against these instances, because of a lack of control over the system that has significant control over you. Therefore, I believe it is exciting to see whether we can argue that these systems are dominating us. A possible form of resistance against this presumed domination should incorporate the worries of accountability and transparency that were sketched in this chapter.

III. Domination

When we think about domination, the accompanying thoughts most likely involve a kind of relationship between the state and the commoner, restricting freedom and imposing their ideas without the possibility to influence those ideas (Pettit 1996). It is based on an asymmetrical power relationship between two parties. The core of this thesis is based on algorithmic decision-making that affects our daily lives to quite an extent. In the previous chapter, we discussed some accountability issues in those decision-making processes. What does it mean that we are heavily influenced by decision procedures using algorithms that we cannot hold accountable like we can hold human beings accountable? This chapter will focus on whether this asymmetrical relationship could be considered a form of domination.

3.1 'If the service is free, you are the product' – the concerns of Big Data

The power of algorithms is determined by the quality of coding, the machine learning involved, and the quality (and amount) of input: data.

When considering datasets that increasingly influence our lives, we have to look at the concept of Big Data. According to Bernard Marr, Big Data entails that we can collect and analyze data in ways that were not possible before – we have more of it and are more capable of storing and analyzing it in light of our interests (2016, 1). We are 'producing' this data every time we use anything related to the digital world, varying from sending an email or WhatsApp, the GPS sensor in our mobile devices, and our interactions with social media by liking, posting, tweeting, and even just scrolling through our feeds. With the right algorithmic tools and data analysis, companies that use Big Data can adjust their products to your interest, personalize advertisements to increase their profit, and improve overall performance. With a thorough analysis of data, algorithms can even predict terrorist acts by combining data and detecting patterns impossible to detect by a simple human being (2016, 3). In his book 'Big Data in Practice', Marr describes several Big Data successes and how major companies like Apple, Google, and Facebook use Big Data to adjust their business to please you while staying in line with their interests.

Big Data is in the hands of large companies, meaning that the primary sources of influence over our behavior and a lot of our data are not in the hands of the government. Those companies can have different intentions regarding what to do with these extensive data sets. Either way, data is incredibly valuable to those companies. Multinationals like Google and Facebook offer their products for free, but in fact, you pay them with your data. The famous expression 'if the service is free, you are the product' represents this business[4]. When a notorious product is offered to you for free, it often implies that the input you give them is far more valuable to them than a small fee from all of those would be. For example, in the case of Facebook, your data could be used to tailor their advertisements, earning them far more money than a fee from your side would earn them. The far-reaching possibilities with data make your personal information more valuable than gold.

This all is not necessarily a bad thing. The use of Big Data has provided us with the tools to predict things beneficial for us – consider the terrorism trait as an extreme variation of what useful Big Data can do, and consider Netflix recommending you an unknown movie that you end up enjoying as a lighter example of the benefits of Big Data. However, this form of '*automated aim attribution*' can also be considered ethically problematic (King 2020). We should also consider the possible side-effects, especially in light of our accountability issues in algorithms, as discussed in the previous chapter. What does it mean that companies own those Big Data sets which they can use for their good?

According to Frank Pasquale, authority is increasingly expressed algorithmically (2015, 8). He also argues that automatically generated decisions can improve the quality of our daily lives, but asks himself where we will call a halt. Related to the concern of multinationals and Big Data, he argues: '*The contemporary world more closely resembles a one-way mirror. Important corporate actors have unprecedented knowledge of the minutiae of our daily lives, while we know little to nothing about how they use this*

[4] The phrase appears in many different forms, of which this is the core message. The quote gained popularity when written down as a comment by user 'blue_beetle' (identified as Andrew Lewis) on the blog 'MetaFilter' in 2010 (Lewis 2010)

knowledge to influence the important decisions that we – and they – make' (2015, 9). Furthermore, he argues that we are mostly ignorant of how these multinationals interact and conflict with public powers, citing Jeff Connaughton by arguing that we are 'increasingly ruled by "the Blob", a shadowy network of actors who mobilize money and media for private gain' (2015,10).

In short, not only the state but also companies can display authority over us through their algorithm usage. A possible follow-up question, and the main focus of this thesis, could be whether or not those algorithms are – in some way – 'dominating' us. John Danaher argues that algorithmic tools can enable distinctive forms of domination, which he calls 'micro-domination' (2019, 108). To provide a possible answer to this question, we must consider the concept of domination first. The influential philosopher Michel Foucault provides us with the best-known conception of what domination entails.

3.2 Foucault's microphysics of power and normalization

The first pages of his influential book on the birth of the prison 'Discipline and Punish' describe the public display of the execution of Robert-François Damiens the regicide on March 2, 1757 (Foucault and Sheridan 2012, 3). This is what Foucault calls sovereign power: power as physical punishment, and punishment as a public spectacle to display sovereign power (Kallman and Dini 2013, 139). During these days, prisons were only used as a punishment in particular cases. The sovereign displays his or her power over the subject with force, and according to Foucault, we should not regard these punishments solely as a way to punish the subject for his or her deeds, but also as a 'complex social function' and a 'political tactic to change behavior' (2012, 23). This change in view on power relations and their purpose explains the increased popularity of the prison: not only is it used to punish the ones who broke the law, but also as a way to correct their behavior to ensure that they will not display the same behavior in the future. This change in view indicates that we need a new theory of power to understand how these power

relations work and is the main aim of Foucault's theory of disciplinary power[5] and subjection. He argues that subjection to power is not only achieved by violence, but rather by a complex 'microphysics of power' that disciplines more subtle, is exercised rather than possessed, and exists not in individuals, but in relations – and on every societal level (Kallman and Dini 2013, 141).

This 'microphysics of power' refers to the techniques by which discipline invests power in the body. It is a technique to maximize individual efficiency, to let power not only impact you negatively (by punishment) but also positively by producing new behavior. An important notion here is what Foucault refers to as 'normalization': to improve the efficiency of the individual, we must not look at the individual judicially, but strategically. This means that we do not judge individuals based on their wrongdoings or virtues in light of the text of the law, but in terms of their behavior in relation to the general norm in society. Foucault writes: '*The perpetual penalties that traverse all points and supervise every instant in the disciplinary institutions compares, differentiates, hierarchizes, homogenizes, excludes. In short, it normalizes*' (2012, 183). Foucault also refers to 'governmentality' to conceptualize these power relations, focusing on the 'way in which one conducts the conduct of men', and we can analyze discipline, biopower, and normalization in terms of governmentality (Simons et al. 2013, 311).

One of the most prominent displays of this new 'microphysics of power' is Jeremy Bentham's 'panopticon' (1791). The panopticon's architecture displays efficiency at its finest: a circular building with a watchtower in the center, designed in such a way that a possible guard in the tower can see all inmates without them knowing when they are being watched. The efficiency in this design is optimized because the microphysics of power ensures that the same result is reached when no one is in the tower as when someone is in the tower – because the inmates are unaware whether they are being watched or not, they start behaving in accordance to the will of the guard just to be sure.

[5] In some cases, Foucault also similarly speaks of 'biopower'. Biopower and disciplinary power are mostly complementary but can be used for different purposes, where biopower is more suitable for discussions on, for example, sexuality, birth control, and mortality rates (Simons et al. 2013, 305).

As Foucault puts it: the major effect of the panopticon is '*to induce the inmate a state of conscious and permanent visibility that assures the automatic functioning of power'* (2012, 201).

Disciplinary power is everywhere, and everyone is subjected to it, even those who exercise it (Allen et al. 2013, 345). What does this subjection entail? Allen et al. describe that for Foucault: 'the individual is an effect of power, but the individual is also always the 'relay' of or conduit for the power relations that make her who she is' (2013, 346). Although Foucault famously argues that where there is power, there is resistance (Foucault 1978, 95), Allen et al. argue that the notions of resistance and subjection are underdeveloped (2013, 346). We can resist to the changing forces of power, but how one should do so remains unclear.

Foucault describes a situation in which we, as human beings, can be dominated by disciplinary power relations that exist between us and other human beings on every societal level. However, we must ask ourselves how this concept can be applied to algorithmic decision-making specifically. It poses the question of whether Foucault's conception of domination applies to algorithms. A relatively new field in which Foucault is often quoted in the relation between new technologies and domination is in surveillance studies. Therefore, it is helpful to see how they see the relationship between new technologies and domination in a Foucauldian sense to see how we can apply this concept to algorithms specifically.

3.3 Domination in surveillance studies

In recent articles, much attention has been given to a Foucauldian (and Deleuzian) concept of control in surveillance studies (Wood 2007; Vanolo 2014; Hoye and Monaghan 2018; Sadowski and Pasquale 2015; Haggerty and Ericson 2000). Deleuze, mainly inspired by the concept of disciplinary power as put forward by Foucault, argues that the *societies of control* are in the process of replacing the system of disciplinary power structures (1992, 4). The focus of societal control is no longer the individual, but merely the 'dividual' – we are regarded as pieces of data and passwords, free to go

wherever they want to go as long as their passwords seem to work (1992, 5). Surveillance can ensure that the suspects' behavior is produced by the effect of the power-relations that it is subjected to. If we combine this with Foucault's notion that this disciplinary power is inescapable and always present, this raises questions concerning the freedom of citizens in our new 'surveillant societies' – societies in which state surveillance becomes the norm and individuals are more than eager to give away their personal information to large privacy-invasive companies like Facebook and Google, who use the information of *dividuals* to present highly specific recommendations and advertisements. For their algorithms, I am not Kelly the individual, but an extensive collection of interests. It is not a coincidence that Facebook offers incredibly elaborate targeting options, which vary from basic information like your gender and age to tremendously detailed information like your previous charitable donations, whether you are likely to move and your buying habits on websites that are not directly linked to Facebook in the first place (Lister 2020). We leave a digital footprint every step along with our digital ways. De Laat describes the combination of these different contexts as a *'polypanopticon'*, in which the '*panoptic gazes of many different contexts are coupled together*' (2019, 323). This combination of data from different contexts may be surprising. Using smart algorithms, my digital traces from online shopping and my social media habits on Facebook may impact any future online application that deals with smart algorithms. An effect of which I may be completely unaware.

Hoye and Monaghan argue that responses to this form of power have been ineffective. In essence, liberal critiques are constrained by their conception of freedom as non-interference, and Foucauldian critiques of liberalism have been incapable of addressing the normative questions regarding the relationship between surveillance and freedom (2018, 343). Their focus is on the neo-republican idea of freedom as non-domination, derived from Pettit (1996; 2002; 2012; 2014) and Skinner (2004; 2010; 2012). This is understood as 'being subject to the arbitrary will of another agent irrespectively of whether or not the dominating agent interferes. Unfreedom consists of the agent's status as dependent upon another's power' (Hoye and Monaghan 2018, 348). Therefore, a free agent is not subjected to the will of other agents that can interfere and has control over those powers that do interfere (2018, 348). This neo-republican view agrees with the

Foucauldian concept of control and power as described to a certain degree: the analysis of governmentality, the notion that power can shape behavior through non-invasive means and regard domination as a 'non-interfering power that constrains by way of power relationships and corresponding (dis)incentives' (2018, 350). It is in a sense a form of nudging by design, in which the exercise of power is not intended to restrict the choices of the individual, but 'directing' the individual's behavior in such a way that the preferred outcome is reached. According to Hoye and Monaghan, domination happens when governance does not track citizens' interests in these instances, so when democratic inputs are lacking (2018, 350).

Foucault is especially influential in the understanding of surveillance as a governing technology and turning negative surveillance into productive surveillance. However, he does not provide us with tools to distinguish agent-ascribable domination from agentless domination (Hoye and Monaghan 2018, 350). This agentless domination is what makes surveillance such a distinct case, making individual agents inconsequential. Whether the agent is ascribable or not does not necessarily make a difference if the problem is that the individual has the fear or intuition that it is being dominated (Kateb 2006, 99). The problem of domination in light of surveillance is not that the state or large companies are actually tracking our daily conduct; it is more specifically the fact that they have *the power to do so*.

De Laat emphasizes the similarities between Foucault and the use of algorithms in decision-making processes by arguing that predictive modeling can be interpreted as a Foucauldian discipline with a twist (2019, 323). He stresses the role of *normation* in these algorithms, relating to the concept of 'normalization' caused by the workings of the microphysics of power. Recommendation algorithms using machine learning use their input from society to increase the algorithm's efficiency and success. Within the data set that society delivers, certain norms are already incorporated, and underrepresented groups in society are most likely also underrepresented in those data sets. In this way, the procedure of recommendation systems ensures that the normation as present in society is extended (2019, 323).

These views can be of great importance to whether surveillance can be a form of domination using Foucauldian concepts. However, algorithmic decision-making is not the same as mass surveillance. Therefore, a critical question regarding the possibility of domination by algorithms specifically is not directly addressed in this literature. 'Surveillance' can be seen as a political system that, in light of Foucault, can govern our behavior using the microphysics of power. That does not directly entail that *algorithms* can display this behavior. The question arises: can 'things' actually dominate? Or do they require some form of agency to do so? More more importantly, if we argue that they can, what does that imply for the status (and, related, the question of accountability) of machine learning algorithms?

3.4 Can 'things' dominate?

Danaher argues that what he calls 'algorithmic tools' can enable distinctive forms of domination. Just like Hoye and Monaghan, Danaher builds his argumentation on the neo-republican concept of freedom as non-domination. He writes: '*We are all watched over by algorithmic tools of loving grace, each of which is standing in wait to nudge us back on track if we ever try to escape'* (2019, 108). He uses Tom O'Shea (2018) to introduce the term 'algorithmic micro-domination'. O'Shea argues that domination consists in arbitrary power, and '*an agent dominates another to the extent that they have the capacity to interfere, on an arbitrary basis, in certain choices that the other is in a position to make'* (2018, 134). The ability to interfere – just like in the panopticon – can be enough to subject people to domination. O'Shea introduces the term 'micro-domination' to describe the capacity for decisions to be arbitrarily imposed on someone, specifically focusing on the concept of domination experienced by people with disabilities (2018, 136). He specifically uses the term 'micro' because the cases of domination are often too minor to be handled in court but have a significant impact on the lives of those affected[6].

[6] The idea of 'micro-domination' may be similar to the concept of 'microagressions'. Microagressions are relatively minor insulting events and indignities who are very harmful because they are a part of an oppressive pattern of similar insults (Rini 2018). While the use of 'micro' is overlapping (events that seem *minor* but have a *significant harmful impact* on the one subjected by it), micro-domination differs from microagressions in their focus. Micro-domination primarily focuses on the power to actively interfere with arbitrary power in the set of choices of the subjected, administering their lives (O'Shea 2018, 137). It involves making choices on behalf of the one subjected, like forced treatments for human beings who have a cognitive disability. The literature on microagressions does mention the case of humans with disabilities who are routinely insulted and *'whose insults are linked to stable*

Danaher argues that the systematic use of algorithmic tools across various domains can give rise to a similar phenomenon. Many 'minor' choices in our daily lives are influenced by or executed with the help of an algorithmic tool – making us 'subjects' of many algorithmic masters. They would 'surveil our lives and create a space of permissible/acceptable behavior' (2019, 108–9).

Who are those 'algorithmic masters' that Danaher is talking about? Can we even ascribe agency to machine-learning algorithms that are part of a decision-making process? Closely related to the accountability problems in algorithms as discussed before, the term 'algorithmic micro-domination' poses the question of *who* or *what* is precisely dominating in these cases. The question of agency concerning algorithms is, at this point, mostly present in debates on responsibility and automated systems of which machine-learning algorithms are part, for example, in automated driving systems and autonomous weapon systems (see for example Di Nucci and Santoni de Sio 2014; Ekelhof 2019).

According to Nyholm, making decisions and choices are key aspects of our ordinary conception of agency, and many authors writing about autonomous systems attribute a significant and highly autonomous kind of agency to these systems (2018, 1204). autonomous systems will work independently from direct input from human agents; therefore, they will make the decisions. Consequently, it would be unfair to hold any human agents responsible for the robotic agency exercised by the automated systems. However, it is a far stretch to argue that these systems work independently from human beings, so far that we can attribute the same kind of agency to them as to human beings. In my example of the algorithm determining whether I get a job or not, the system is in some way designed by a human being – at least, its purpose is created by the company that uses the system to hire people, and the input and feedback to the system (that

traits such as gender, ethnicity, or disability status' (Rini 2018, 334). However, in these cases, while one could argue that being oppressed by a system can influence your choices in your daily life, microagressions are not necessarily the consequence of asymmetrical power relations or arbitrary decisions made on their behalf. Therefore, within this thesis, the phenomenon of microagressions is interesting because it displays the harmfulness of relatively minor expressions, which can have great impact on the lives of the ones affected by them, but they are not per se an example of domination in my conception of the term.

ensures the machine learning functionality of the system is useful) are also done by human beings. In a sense, the system is designed by human beings.

Nyholm introduces four different agency types to demonstrate what kind of agency we can attribute to autonomous systems and what types we cannot (2018, 1207–8). The four types he introduces are:

1. Domain-specific basic agency
2. Domain-specific principled agency
3. Domain-specific supervised and deferential principled agency
4. Domain-specific responsible agency

He demonstrates the limits of attributing agency to autonomous systems by applying these four types of agency to the case of autonomous driving systems. The first three can all be attributed to autonomous driving systems. A self-driving car can drive to its destination in a way that is sensitive to representations of its environment (1), is programmed to follow traffic rules (2), and is being watched over by an authority who can interfere and to whom the car is deferential, e.g., the person in the car or the engineers who design the software (3). However, the fourth type of agency that focuses on the system having the ability to understand criticism and defend one's actions cannot be attributed to the system. Therefore, the system cannot take responsibility for its actions like human beings can (2018, 1209). This can be applied in a similar way to the algorithms discussed in this thesis. The algorithms incorporated in the hiring system of the company recommend who to hire (1) and can do this following a set of rules put forward by the company in the initial design of the algorithm (2). Furthermore, if the company is not pleased with the algorithm's output, it can interfere with the algorithm and either stop using it to make recommendations or change the set of rules the algorithms adhere to achieve the desired outcome (3). However, just like in the case of the self-driving car, the recommendation algorithm cannot actively discuss with me or the company about the decisions that are made or can give any kind of justification for the process apart from the rules initially provided by the company. This process is even

more challenging when we are concerned with machine-learning algorithms because the algorithm's machine-learning capabilities are opaque.

Furthermore, as Nyholm argues, the agency performed by the system are mostly in response to someone else's initiative (2018, 1211). This makes the system's agency of a collaborative type. This can also be applied to our example. It is not as if the recommendation algorithm has any intention to judge input on how well they would fit within a company. For the algorithmic system to make any sense, it needs to be collaborative with a company that is interested in hiring new people and prefers an algorithmic system to perform this task over themselves because of their reasons, like efficiency and accuracy. Using Nyholm's argumentation, we can argue that algorithmic systems do demonstrate a form of agency, but that this form of agency should not be considered entirely autonomous and independent from human agents. Therefore, when we are concerned with responsibility issues, we should look at the collaborative agency at stake and find the human beings that are 'most' responsible for the system's behavior.

However, in practice, the allocation of responsible human agents can remain unclear or can be very difficult. Roos de Jong argues that worries about responsibility gaps are still justified because it often remains unclear how to distribute responsibility among the relevant human agents involved (2020, 727). Especially when multiple groups and individuals attribute to the workings of the system while not having an evident collaboration with each other, it can be difficult to trace who is exactly responsible for what part of the system's behavior (2020, 731). So, while the concept of collaborative agency is of great use for this thesis, the practical implementation can be troubling. Therefore, this issue will be addressed in the next chapter of this thesis on resistance to investigate whether a possible form of resistance against algorithmic domination could make the attribution of responsibility to involved human agents clear by overcoming De Jong's worries.

Using the concept of collaborative agency, we can pose the question of whether the algorithmic system itself can dominate us, or if we have to expand this view to the human agents involved in the domain-specific supervised and deferential principles

agency. Is it fair to attribute the label 'dominator' to the algorithmic tools included in the company's decision-making process, or should we pay closer attention to the human agents involved in the design of the tools?

On the other side of the agency debate are authors who argue that we do not necessarily need to acknowledge the dominating power as an agent for them to dominate (or oppress) us. Liao and Huebner, for example, argue that apart from psychological and social components, physical components should also be included in the analysis of oppressive systems like racism and sexism (forthcoming, 2). This entails that 'things' can also be oppressive. Ascribing oppressiveness to a physical object can be of interest to this thesis because it could indicate that we could, perhaps, also attribute domination to an algorithm irrespective of the involvement and division of agency. However, their case cannot be directly applied to the case of algorithms. First of all, their focus is primarily on oppression like racism by giving excellent examples like Kodak's Shirley card, which was used by photographers to calibrate skin-color balance during the printing process. However, Liao and Huebner describe an exciting phenomenon in oppression relevant to our understanding of domination. The authors describe the central issue of oppressive things like the Shirley card: it provides a '*prescriptive standard against which variations are treated as deviations from the norm'* (forthcoming, 4). This argument closely resembles the concept of normalization from Foucault, and a consequence of the workings of the microphysics of power.

How can Liao and Huebner make the argument that a physical object can be oppressive? In their argumentation, they describe that a thing can be oppressive when it partially constitutes the stability and structure of a specific oppressive framework (forthcoming, 7). They specifically use Langdon Winner's political framework to argue that artifacts can embody political systems insofar as they are "convenient means of establishing patterns of power and authority in a given setting" or have "inscrutable properties that are strongly, perhaps unavoidably, linked to particular institutionalized patterns of power and authority (Winner 1980, 122, 134). Extending Winner's view, Liao and Huebner's framework holds that things are racist when they are *congruent* with an oppressive system such as racism. At the end of their article, they mention the parallel

between their view on oppressive things and biases within algorithms, arguing that "*philosophical investigations of algorithmic bias demand a framework that is more sensitive to oppressive systems at the start"* (forthcoming, 16). Similarly, Safiya Umoja Noble argues for the existence of 'algorithmic oppression', in which the algorithmic data failures in Google's search engine show congruence with the oppressive frameworks of racism and sexism (2018, 4).

An important side note here is that I do not believe that the case that Liao and Huebner make with oppressive things can be directly applied to the matter of algorithmic domination. There are two reasons why I think this is the case. First, Liao and Huebner focus on 'material artifacts.' As I also put forward in chapter one, there is no clear definition of what an algorithm or an algorithmic system is. Furthermore, especially smart algorithms are continuously changing due to their machine learning capabilities. Second, oppression and domination are not necessarily the same 'things' – at least, in my view, they are different in their primary focus. Domination, at least in this thesis, primarily focuses on a Foucauldian concept of domination, whose focus is on asymmetrical power relations. The idea of domination in surveillance studies also focuses on the 'ability to interfere' and our 'powerlessness' in these situations. Domination is a consequence of the active and productive workings of the microphysics of power. On the other hand, oppression in this context focuses on institutionalized oppressive systems like racism and sexism, that are being maintained in physical components. The issue presented here has some similarities in the algorithmic bias debate, which is incredibly interesting and important for further research. While algorithmic bias and normalization, as put forward by Foucault, are undoubtedly interesting for domination, it is not the debate that I wish to address in detail in this thesis.

However, their extended version of Winner's framework could be of interest to the question of whether algorithms are capable of domination. Imagine that we argue that things can be a dominating agent when they are *congruent* with a dominating system. To be congruent, the 'thing' should be biased in the same direction as the oppressive

system, be causally embedded in the oppressive system, and this connection must be bi-directional – ensuring that it actively guides and constrains psychological and social structures (forthcoming, 10).

It is difficult to attribute these capabilities to algorithms alone within a recommendation system. When we look at the first precondition, for example, agency problems already arise. There are various cases in which technological systems, such as face recognition software, are *biased* in the sense that it tends to work noticeably worse on people of color. In research in influential facial analysis software used by large companies like IBM and Google, it was found that machine learning algorithms discriminate based on class like race and gender by misclassifying darker-skinned females in up to 34.7% of the cases. In comparison, the error rate on white males was found to be 0.8% maximum (Buolamwini and Gebru 2018). However, does this mean that we can argue that the algorithm itself is discriminating? In my view, we cannot. While the research made clear that there is a substantial problem apparent in these algorithms, it demonstrated that the datasets of these facial recognition algorithms were overwhelmingly composed of lighter-skinned subjects. As mentioned before, even machine learning algorithms can be as good as their input, and it makes little sense to blame the algorithms that their input is hugely biased to begin with. So, even within a framework in which it seems possible to attribute oppression to a physical component, it does not make sense to assign domination entirely to an algorithmic system only. At its best, we can argue – using a concept like a collaborative agency – that while the algorithm is doing the job and producing the output, we should look at the bigger and cooperative picture to identify the agents involved and whom we can hold accountable for it. This means that when Danaher speaks of 'algorithmic masters', we should look at the cooperation between the algorithmic system that is used, and the relevant human beings involved in the workings of that system to determine against whom we should resist.

A proposal to avoid the problem of agency is the '*panopticanesque'* argument from Hoye and Monaghan, which states that there is a concept called 'agentless domination'. They write: *'The new power of surveillance functions as though there were no ascribable agents – not because there are no agents within the aggregate, but because the levels of*

imbrication, secrecy, and redundancy are so high as to make individual agents inconsequential' (2018, 351). Again, they do not solve the question of whether we can regard algorithms as agents who display dominating power over human beings. This argument tells us that there is domination, regardless of the agents responsible or involved in the process. The problem with this view is that it is difficult to do something about domination (in a Foucauldian sense: to 'resist'), when you do not know who or what you should be resisting against. Therefore, it would be helpful to identify the relevant actors involved in the process, to be able to provide what Hoye and Monaghan describe as an *anti-power* against the dominating powers at play. The next chapter of this thesis will, therefore, focus on resistance.

IV. Resistance

In the previous chapter, I deliberated on whether algorithms and algorithmic systems dominate us. I concluded that algorithmic tools show symptoms of domination in the Foucauldian sense, as also demonstrated in surveillance studies. Furthermore, research on the concept of agency showed that it is nearly impossible to hold the algorithmic systems themselves accountable for this domination. We should, therefore, look at the cooperation between human agents and the algorithmic systems. When we accept the idea of domination in this context, a logical follow-up question would be: how can we prevent this? This chapter will focus on answering this question. I will use Foucault's and surveillance studies' analysis of resistance to draw the preconditions for a form of resistance. After that, I will propose the method of meaningful human control as a form of resistance against algorithmic domination and give an analysis of whether this proposal is successful in removing the threat of algorithmic domination.

4.1 The anti-power

Foucault famously argues that '*where there is power, there is resistance, and yet, or rather consequently, this resistance is never in a position of exteriority in relation to power*' (Foucault 1978, 95). However, the specific notions of resistance and subjection are underdeveloped (Allen et al. 2013, 346). We are all *able* to resist to the changing forces of power, but *how* one should do so exactly remains a bit unclear. In surveillance studies, this resistance is seen as a sort of anti-power. According to Hoye and Monaghan, a possible way to exercise anti-power would be to monitor the power relations as a starting point and to put boundaries around the dominating power to institutionalize capacities for resistance (2018, 355-356). This involves a sort of counter-surveillance strategy to expose but to do so reliably over variation and time. The republican approach is rarely to eliminate the source, but to put boundaries around it. However, while this solution may sound ideal, it also faces similar problems as our conception of agency does. We need to, therefore, identify the dominating powers involved in the processes.

To construct an anti-power, it can be helpful to identify how technology can exercise power and what their role in that process should be like according to political theory. Philip Brey argues that a critical political theory of technology is a theory that interprets and criticizes the role of technology in the distribution and exercise of power in society (2008, 72). His main aim is to ensure that technology will be a form of empowerment instead of domination. In this way, Brey tries to combine the tradition of critical political philosophy with technology. The political critique of technology found its origin in the works of Karl Marx and the Frankfurt school (e.g., in Feenberg 1996). Other relevant authors on this topic, who I will not discuss in further detail in this thesis but can be of interest, are Winner (e.g., 1977) who focuses on the political critiques of technology, and Latour (1987; 2004) who focuses on power relations between technical artifacts and human beings. A critical analysis of the political critique of algorithmic systems can be of great interest to the debate overall but goes beyond the scope of this thesis. Therefore, this could be an interesting topic for further research.

According to Brey, the individual can be empowered if we develop a better resistance against social power exercised by others (2008, 75). The three virtues that should be strived to maintain a good society are freedom, democracy, and justice (72). Brey specifically argues that we should strive to democratize technologies to ensure that all relevant stakeholders – not only the companies and the engineers but also the users of the technologies – are involved in the design process (92). This emphasis on democracy was also mentioned by Hoye and Monaghan: domination occurs when the system does not track the interests of the citizens involved. Therefore, a form of democracy should be incorporated into the design of the algorithmic systems. Moreover, concerning our posed problems with accountability, and the cooperative agency proposal, part of the resistance against this domination entails that we make parts of the algorithmic system more transparent.

An important counterargument to consider in this thesis is the idea that this all is not new. As Danaher poses in 'the Ethics of Algorithmic Outsourcing in Everyday Life', one could argue that domination by algorithmic tools is just '*an old wolf in new clothing'* (2019, 109). It is not as if I was never dominated by external parties before algorithms

occurred, which makes sense if we follow Foucault's theory on disciplinary power and agree that this power is everywhere. However, it does not really matter for my argument whether or not this form of domination is new, or just a modernized form of what we have been dealing with before algorithmic systems. The issue of domination and the need for a form of resistance stays the same. What has changed, and in this I agree with Danaher's argumentation, is the scope of domination. Algorithms are everywhere online, and their learning capabilities are unprecedented. Therefore, they can dominate you in a far more personalized way than before without you even realizing it, which makes resistance difficult. Resistance should focus on informing the individual subjected of their subjection to make her more aware of her position and the power she can exercise over her dominators if she wishes.

When we apply all this information on our case of decision-making algorithms, some specific preconditions for the design of such systems appear. In short, a possible form of resistance should (1) uncover the asymmetrical power relations involved in the decision-making process, in order to (2) identify the relevant agents involved, (3) incorporate democratic values and track citizens interest to empower them, and (4) make the overall system more transparent to the individual. This does not mean that we can altogether remove the threat of domination or the dominating powers themselves, However, it should give us a sense of control over the processes that substantially influence our daily lives.

These ideas and ideals are not new within the new context of machine-learning autonomous systems. In fact, these preconditions seem to resemble the theory of *meaningful human control* closely. Therefore, it can be interesting to see whether this existing theory can be of use in our scenario. I will summarize the relevant aspects of the theory of meaningful human control and analyze whether this existing proposal can be of help to resist our case of algorithmic domination.

4.2 The theory of Meaningful Human Control

This thesis will focus on the theory of meaningful human control as put forward by van den Hoven and Santoni de Sio, and improved and extended by Santoni de Sio and Mecacci. Their view is concentrated on the case of automated driving systems (Santoni de Sio and van den Hoven 2018; Mecacci and Santoni de Sio 2019). Mecacci and Santoni de Sio use the example of dual-mode vehicles like Tesla's autopilot to make their argument: a specific case in which the human driver gives up part of the control over an action by delegating that part to an autonomous system (2019). The aim is to investigate which kind of control is required to maintain high levels of *safety* and *accountability* (Sparrow & Howard, 2017, quoted in 2019). This focuses on the kind and extent of control that human beings have over an automated driving system. One could say, perhaps metaphorically, that if we do not have this kind of control over the autonomous system, the technology takes over and could dominate human beings in general.

According to the Santoni de Sio and Mecacci, a concept like meaningful human control is necessary to address possible responsibility gaps and ensure safety by promoting a stronger and clearer connection between human agents and intelligent systems (2019, 3). This theory is based on the theory of *guidance control* as provided by Fischer and Ravizza (1998). Guidance control is realized when the decisional mechanism is "moderately reason-responsive" and is "the agent's own" (1998; quoted in Santoni de Sio and van den Hoven 2018). In meaningful human control, using Nozick's theory of knowledge, these two conditions are called *tracking* and *tracing* (2018,6).

The tracking condition holds that there needs to be a *'tracking relation between human moral capacities to respond to relevant moral reasons and (military) system actions'* (Santoni de Sio and van den Hoven 2018, 6). Therefore, this condition ensures that human reasons will always be integrated in the workings of the system. The tracing condition argues that the actions of the system should be traceable to *'a proper moral understanding on the part of one or more relevant human persons who design or interact with the system',* ensuring that there is always at least one human agent who

understands the system and its consequences and is also aware of how those affected by the system may respond to the actions of the system (2018, 9).

Meaningful human control focuses primarily on control in terms of a relationship between human intentions in general and autonomous systems. In that sense, we can regard meaningful human control as a form of resistance against the possible domination of the autonomous system. However, there is another way to regard meaningful human control as a form of resistance. This requires a focus on the individual feeling overwhelmed by the autonomous system because it does not respond to the individual's intentions explicitly. Meaningful human control as it is argued for in the literature mentioned does not explicitly address the possible conflicts of power between the relevant agents, but using a Foucauldian concept of control, the neo-republican view on freedom and Brey's values, I believe that meaningful human control can address these situations, and act as an anti-power against political and societal domination. To argue for this option, we first need to take a closer look at the tracking condition.

The authors emphasize that their definition of tracking does not specify whose relevant reasons should be tracked, only that they should be human. On top of this, it is not necessarily the case that the system is solely influenced by good goals or values. (2018,8). Mecacci and Santoni de Sio establish a reference framework to represent how different reasons stand in reciprocal relation and how they stand in relation to a system's behavior (2019). The proximity scale presents how distal and proximal reasons and agents influence the behavior of the system. This scale shows that 'society' is an agent, though distal, and general societal norms can be seen as 'relevant' reasons the system should adjust to. Since the tracking of relevant moral reasons can also apply to other agents than the driver, we need to take into account a more diverse number of potential agents involved in control tasks. According to the authors, these agents can all be potential controllers of the driving systems insofar as their reasons are reflected in the system (2019).

Multiple agents are involved in the situation of a driver in an automated driving system. Behind the automated driving system is a team of policymakers, designers,

programmers, and other experts who want to influence the system as much as possible. Therefore, it is possible that these intentions vary from and conflict with each other while all have a certain impact on the system's behavior. As an individual without meaningful human control, those influences are invisible to you, while they do influence the behavior of the system you just gave up your control to. This is where Hoye and Monaghan come in again: their definition of free agents is that one should not be subjected to the will of other agents that have the capacity of interfering, and one should have control over those powers that do interfere. In this case, neither of those conditions is fulfilled. In the automated driving system, the individual is subjected to the will and influence of other agents than himself. Since those influences are invisible to you, you have in no sense control over those interfering powers.

The tracking condition can provide us with the tools to map these power relations. The classification of relevant reasons can correspond to the mapping of power relations. In light of disciplinary power, it is necessary to see which actors are involved and to whose relevant reasons the system should respond. It is reasonable (and desirable) that the system should respond to agents like policymakers because technology needs to be safe and responsible. However, these power relations should not become too asymmetrical to prevent the domination of the individual. The theory of meaningful human control is not only necessary to investigate who is in charge and who is responsible when things go right or wrong, but also to investigate whether there are asymmetrical relations of power involved and what the influence of other parties than the agent himself is. Other parties include the programmers of the system, the companies involved, and the role of the state.

A possible way to use meaningful human control to prevent these asymmetrical power relations from rising is to democratize the technology. This is one of the values Brey mentioned, and just like Hoye and Monaghan argued for when they stated that the interests of citizens need to be tracked in order to prevent domination. It is not enough to merely state that the individual has the opportunity to interfere with the system, e.g., by braking or speeding when it wants to, if he does not know to whose interests the system is designed and lacks the knowledge to interfere in a substantial matter. In a

sense: we cannot expect individuals to resist against asymmetrical power relations if they do not know about these relations, let alone how and to which extent the individual can influence the system in the first place. By democratizing the technology, we can empower the individual, for example, by involving her in the design process and/or policymaking for the automated driving system, education, and the provision of manuals.

Furthermore, by democratizing technology and the process that belongs to it, we make agentless domination practically impossible. This is the case, not only because we make transparent whose reasons have what kind of influence on the system, but most importantly because we prevent the individual driver from being dominated in the first place. Meaningful human control can thus not only ensure that humans will not be dominated by the autonomous driving system but can also empower the individual by not letting them get dominated by other human agents in society.

4.3 The application of Meaningful Human Control on algorithmic systems

Meaningful human control seems to be a promising attempt to solve some pressing accountability and domination issues as present within automated systems like driving systems and weapon systems. The question that remains, then, is whether this theory could also be helpful for our issues with algorithmic decision-making.

Algorithmic decision-making in the form that I discussed in this thesis is relatively different from autonomous systems like self-driving cars. As explained in the introduction, I aimed to focus on algorithms that seem smaller in scale and not directly linked to (potentially) lethal situations. The literature on responsibility gaps in autonomous systems focuses primarily on the problem of autonomous systems deciding whom to kill without any form of human control over those systems. However, despite their difference in scope, the central problem remains more or less the same: an autonomous and opaque system displays a form of disciplinary power over you as a user,

of which the details of the workings of the algorithm are unknown to you. Furthermore, the agents involved in the system are unidentified. Can meaningful human control be of use in our case of algorithmic micro-domination? My answer is yes, but in an alternative form.

In the previous chapter, I identified four preconditions that resistance should cover. A possible form of resistance should (1) uncover the asymmetrical power relations involved in the decision-making process, in order to (2) identify the relevant agents involved, (3) incorporate democratic and societal values and track citizen's interests in order to empower them, and (4) make more transparent what the actual decision-making process looks like. By doing so, the form of resistance should prevent 'dominating' agents from obfuscating their responsibility or laundering their agency. Resistance should give the one who is subjected to the dominating power the ability to exercise some form of power and control over the dominator. This would, ultimately, ensure that the algorithmic systems empower the individual instead of merely dominating them.

The theory of meaningful human control (MHC) is useful for our case of algorithmic micro-domination in several ways. The tracking condition of MHC can be used to identify the asymmetrical power relations involved and ensure that relevant agents cannot obfuscate their responsibility by hiding behind the automated system, laundering their agency. The tracking condition can help overcome the problem of opacity so some extent. MHC requires the automated system to track relevant moral reasons and be responsive to them. As argued before, the transparency ideal has significant negative implications, and simply making the system 'more transparent' is therefore not helpful in overcoming domination. MHC ensures that the system responds to relevant moral reasons when necessary. Therefore, the subject of domination does not have to specifically understand the workings of the algorithm on a technical level. MHC's prerequisite states that the system should be responsive to your interests, ensuring that you have some kind of influence and control over the system. In order to do so, I do not need to make the black box completely transparent.

Furthermore, MHC's tracing condition focuses on ensuring that the technology is a form of empowerment of the individual instead of a dominating force, requiring a proper moral understanding of the system. When applied to algorithms, MHC can educate individuals on the workings of system and the relevant reasons involved in the decision-making process, diminishing the feeling of powerlessness.

However, the overall aim of MHC seems to be different than my aim of resistance is within the context of micro-domination. It is useful to acknowledge that meaningful human control over automated driving systems has the overall aim to ensure that there is meaningful human control in *general* over an autonomous system. This means that we can hold a (group of) human being(s) accountable for the workings of the system. However, the domination and resistance I sketched seem to focus more on having control over the system as an *individual*. Does it empower me as an individual if I know that some human being at Google has control over the system if my control over that system is negligible? This difference makes sense because our form of resistance should address not only the automated system but also the human agents involved in their collaborative agency.

Most of these advantages and workings of MHC depend significantly on our interpretation of 'relevant agents' within the system. MHC can empower the individual if we regard the individual relevant enough to be tracked. At this point, MHC faces a serious problem: as it is right now, MHC is still not able to address our most pressing issue regarding agency. As stated above, the tracking condition of meaningful human control states that the system has to respond to relevant human reasons but does not explicate whose reasons that *should* be and to what extent these reasons *should* influence the system. When returning to de Jong's main argument against Nyholm's form of cooperative agency, arguing that it is often not clear how responsibility should be distributed amongst human agents who are to some extent involved in the system, we see that this important problem is not solved yet by a proposal like MHC. It is clear that the relevant agents should be identified, and that in case of sophisticated algorithms and machine-learning systems, it is unfair to hold a single individual accountable when

more agents are involved. However, the precise division of accountability and agency within the pool of relevant agents remains unsolved.

Furthermore, it is the question of what *meaningful* actually means in the case of MHC over algorithmic systems, and from whose perspective this control should be considered meaningful. It is very useful to determine within a context of self-driving cars who should be held accountable for the damage done, especially if such cars are involved in a fatal car accident. And in the case of algorithmic decision-making, it is promising that MHC will attempt to realize that I will have some group, company, or individual to hold accountable and that the system itself is under human control. Nevertheless, in a realistic sense, micro-domination by algorithms is primarily focused on the domination of the individual. The question is: how much control can an individual actually have over multinational companies like Google and Facebook?

I propose to pick out the most useful aspects of MHC for our case of micro-domination and argue that the possible implementation of MHC within algorithmic decision-making could also result in a minor enhancement of the process overall. This is what I would like to call a form of 'micro-resistance'[7]. An important part of why individuals feel dominated by such systems is that they are unaware of how they are subjected to them and how they work. MHC could inform people of the use of algorithms and use it justify the choices companies have made within the design process, and to explain to the individual how they can exercise their MHC over the system. For example, a company could explain choices the system will make, when they will interfere with the system, and what they do to improve their algorithms in case of unwanted outcomes. Showing the individual that the responsible agents actually care about the outcome and the workings of the algorithm could already make a significant impact in empowering the individual in these contexts.

[7] Unsurprisingly, the term 'micro-resistance' is already present within the context of microagressions (Dush 2016). We use the term in a very similar way: daily and relatively minor efforts to challenge the norm, or in my case, the dominating power.

The panopticon showed us how disciplinary power works even when you are uncertain whether you are being watched. The feeling of being watched is enough to adjust your behavior in accordance with the required norm. Therefore, attempting to map the relevant agents, the involved power relations, and equipping individuals with the right and understandable knowledge about these systems, could be as if you suddenly gain the ability to look inside of the watchtower. Micro-resistance could severely impact the workings of disciplinary power in this way when structurally exercised, by focusing on empowerment of the individual.

To summarize, I believe that MHC has great potential in automated driving systems, and certainly is a promising theory in our context of algorithmic decision-systems. The theory focuses on the right problematic aspects of opaque algorithms: the possibility of responsibility gaps, the problem of transparency, the identification of relevant human agents involved in the design, and workings of the system and the workings of (disciplinary) power within the system. However, it is unable to answer our most pressing issue: who *are* the relevant agents involved and how do we *distribute* responsibility amongst them? Therefore, MHC's impact as a form of resistance is somewhat limited to the ideal of identifying relevant agents, the prerequisite that our interests should be tracked by the system, and the importance of a proper understanding of what is going on. While this seems minor, I do believe that it is a great step in the right direction. Resistance's aim, according to our neo-republican framework, was never to actually diminish the disciplinary power and dominators completely. Resistance is meant to put boundaries against the dominating powers. And while the empowerment of the individual may be limited, I believe that we can call MHC a form of micro-resistance. And following the other micro-terms mentioned in this thesis, while its content may seem small, it may have a great impact on those algorithmic micro-dominating powers when structurally exercised.

Conclusion

We are confronted with algorithms and algorithmic decision-making systems daily, and they have an increasing influence on our lives. When combined with the immense size and impact of Big Data, they can determine what advertisements I see online and whether I get invited to a job interview. Meanwhile, the system is often opaque, meaning that I have very limited insight into the system that knows everything about me. The research question that this thesis aimed to answer was twofold: are we, individuals who are subjected to opaque decision-making algorithms, being dominated by those algorithms, and if we are, what should resistance against this domination look like?

I have argued that we are indeed dominated by the algorithmic decision-making systems. Arguing from a socio-technological context, algorithms have the capacity to mediate social processes, and regulate and govern our behavior (Mittelstadt et al. 2016; Danaher 2019). To argue that we are dominated, I focused on accountability issues, which mainly revolve around Pasquale's Black Box society (2015), deliberate obfuscation of responsibility, and the transparency ideal. The focus on accountability showed that the opacity of the system and the deliberate obfuscation of responsibility by relevant agents create asymmetrical power relations between the system and the individual affected by the system. I used Danaher's and O'Shea's concept of micro-domination and Foucault's analysis of disciplinary power to argue why this asymmetry in power creates a *panopticonesque* situation in which the dominating parties know everything about the individual subjected to their power, while the subjects know little to nothing about their dominators. The focus is explicitly on micro-domination, understanding domination as the capacity to interfere on an arbitrary basis in choices that the other is in a position to make – and while the cases seem too minor to handle in court, their impact on those affected is significant (O'Shea, 2018). However, I was critical of attributing this dominating power to the algorithmic system itself, wondering whether 'things' can dominate. Using Nyholm's conception of collaborative agency, I argued that we should not see the algorithms in the decision-making system as the dominator, but rather see the bigger picture. While the algorithm is doing the job, we should identify the agents involved to investigate whom we can hold accountable for the domination. Since this

identification of relevant agents faces practical implementation problems (de Jong 2020), I argued that a possible form of resistance should address this issue and establish a framework in which the allocation of relevant agents is achievable.

This brings me to the second question: what should resistance against this domination look like? Resistance should take the form of an anti-power and should empower the individual. This empowerment should encourage the democratization of the system by tracking citizen's interests and by making the process overall more transparent. Furthermore, I argued that resistance should uncover the asymmetrical power relations involved in the decision-making process and to identify the relevant agents involved (and their respective roles) in the process. I deliberated whether the existing theory of meaningful human control over automated driving systems (MHC), as put forward by Santoni de Sio, van den Hoven and Mecacci could be the form of resistance that I was looking for. I argued that while MHC's tracking and tracing conditions are very promising in solving pressing accountability issues within automated driving systems, ensuring that we have a human being to hold accountable, the theory as it is cannot satisfy our desires for resisting against a dominating power from the point of view of the individual. MHC is unable to solve our most pressing issue regarding agency because it does not provide us with the tools to determine whose reasons are relevant enough for the system to respond to. Therefore, in my view, the empowerment of the individual against the system and its collaborative human agents is limited to a minor enhancement: give individuals more information on the overall aim, workings and use of the system, and track their interests as much as possible. I argued that this resistance could be called 'micro-resistance'. While its content may seem small, when structurally exercised, it could have a great impact on algorithmic micro-domination in the end.

I am aware that this research is limited by some of the choices that I have made. I focused on Foucault's interpretation of disciplinary power and domination only. A different interpretation of the concept of domination, or domination in another context than a neo-republican one like proposed, could impact the conclusion of this thesis. Furthermore, I chose to focus on meaningful human control, which focuses on the idea that a human being should always, in the end, be responsible for the workings of an

autonomous system. It is possible to disagree with this, which has a significant impact on the conclusion of this thesis. I also realize that I did not propose a possible answer to the pressing issue of allocation of accountable human agents and the distribution of responsibility among those human agents. While emphasizing the importance of the problem, my proposal of micro-resistance could look like a solution that deliberately tries to avoid solving the problem overall by making its goal less ambitious. However, it was not within the scope of this thesis to give an answer to a question that – in my view - requires a thesis on its own, and I believe that micro-resistance can actually contribute to the debate while we continue to work on the issues that it avoids. One of the aims of this thesis was to emphasize the importance of doing research in the field of machine-learning algorithms that we face on a daily basis. I want to encourage further research on this topic, to emphasize the seriousness of the problem of micro-domination and to encourage the establishment of resistance against those asymmetrical power relations that discipline us subtly.

Further research is needed to investigate what kind of resistance or ethical (and perhaps legal) framework could overcome the problem of identifying the relevant human agents that are involved in autonomous systems. This is not only extremely relevant in our case of algorithmic decision-making systems but is also still a pressing issue within the lethal autonomous systems debate. I encourage further research to incorporate the idea that there is a collaborative agency between the autonomous systems and the relevant human agents and establish a framework of accountability that suits this collaborative relationship. Furthermore, perhaps outside of the scope of the ethics of technology solely, research should be extended outside of philosophy and incorporate a discipline like computer science to investigate how we can make machine learning algorithms in some way more transparent and to see how algorithms can be more democratized in the future. Finally, I would encourage further research of a concept like micro-resistance, which could also be extended to other domains within (political) philosophy. Rome was not built in a day. I am excited to see whether a small change in attitude like micro-resistance could make a difference in a continually changing and technology-craving algorithmic society.

Bibliography

Allen, Amy, Christopher Falzon, Timothy OLeary, and Jana Sawicki. 2013. "Power and the Subject." In *A Companion to Foucault*, 337–52. Chicester, West Sussex, UK: John Wiley & Sons Ltd.

Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media and Society* 20 (3): 973–89. https://doi-org.proxy.library.uu.nl/10.1177/1461444816676645.

Balkin, Jack M. 2018. "The Three Laws of Robotics in the Age of Big Data." *Ohio State Law Journal* 78.

Bentham, Jeremy. 1791. "Panopticon : Or, The Inspection-House." Dublin.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:149–159. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/binns18a.html.

Brey, Philip. 2008. "The Technological Construction of Social Power." *Social Epistemology* 22 (1): 71–95. https://doi.org/10.1080/02691720701773551.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 205395171562251. https://doi.org/10.1177/2053951715622512.

Cohen, J. E. 2012. *Configuring the Networked Self*. New Haven, CT. Yale University Press.

D'Agostino, Marcello, and Massimo Durante. 2018. "Introduction: The Governance of Algorithms." *Philosophy & Technology* 31 (4): 499–505. https://doi.org/10.1007/s13347-018-0337-z.

Danaher, John. 2019. "The Ethics of Algorithmic Outsourcing in Everyday Life," 20.

Deleuze, Gilles. 1992. "Postscript on the Societies of Control." *October* 59: 3–7.

Di Nucci, Ezio, and Filippo Santoni de Sio. 2014. "Who's Afraid of Robots? Fear of Automation and the Ideal of Direct Control." *Roboethics in Film*, 3–7.

Dush, Claire Kamp. 2016. "Fighting Back: Implicit Bias, Micro-Aggressions, and Micro-Resistance." Adventures in Human Development and Family Science. November 11, 2016. https://u.osu.edu/adventuresinhdfs/2016/11/11/implicitbias/.

Ekelhof, Merel. 2019. "Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation." *Global Policy* 10 (3): 343–48. https://doi.org/10.1111/1758-5899.12665.

Feenberg, Andrew. 1996. "Marcuse or Habermas: Two Critiques of Technology." *Inquiry* 39 (1): 45–70. https://doi.org/10.1080/00201749608602407.

Fischer, J., and M Ravizza. 1998. *Responsibility and Control : A Theory of Moral Responsibility*. Cambridge Studies in Philosophy and Law. Cambridge, U.K.: Cambridge University Press.
Fischer, John Martin, and Mark Ravizza. 2012. "Moral Responsibility: The Concept and the Challenges." *Responsibility and Control* 14: 1–27. https://doi.org/10.1017/cbo9780511814594.001.
Foucault, Michel. 1978. *The History of Sexuality*. 1st American ed. New York: Pantheon Books.
Foucault, Michel, and Alan Sheridan. 2012. *Discipline and Punish: The Birth of the Prison*. New York: Vintage Books.
Haggerty, Keven D., and Richard V. Ericson. 2000. "The Surveillant Assemblage." *British Journal of Sociology* 51 (4): 605–622. https://doi.org/10.1080/00071310020015280.
Hart, HLA. 1968. *Punishment and Responsibility; Essays in the Philosophy of Law*. New York: Oxford University Press.
Hill, Kashmir. 2012. "How Target Figured Out a Teen Girl Was Pregnant before Her Father Did." *Forbes* 16.
Hoye, J. Matthew, and Jeffrey Monaghan. 2018. "Surveillance, Freedom and the Republic." *European Journal of Political Theory* 17 (3): 343–363. https://doi.org/10.1177/1474885115608783.
Jong, Roos de. 2020. "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm." *Science and Engineering Ethics* 26 (2): 727–35. https://doi.org/10.1007/s11948-019-00120-4.
Kallman, Meghan, and Rachele Dini. 2013. "Discipline and Punish." In *A Companion to Foucault*, 1–96. https://doi.org/10.4324/9781912282630.
Kateb, G. 2006. *Patriotism and Other Mistakes*. New Haven: Yale University Press.
Kearns, Michael, and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press.
King, Owen C. 2020. "Presumptuous Aim Attribution, Conformity, and the Ethics of Artificial Social Cognition." *Ethics and Information Technology* 22 (1): 25–37. https://doi.org/10.1007/s10676-019-09512-3.
Kraemer, Felicitas, Kees van Overveld, and Martin Peterson. 2011. "Is There an Ethics of Algorithms?" *Ethics and Information Technology* 13 (3): 251–60. https://doi.org/10.1007/s10676-010-9233-7.
Kutz, Christopher. 2004. "Responsibility." In *The Oxford Handbook of Jurisprudence and Philosohpy of Law*, edited by Jules L. Coleman, Kenneth Einar Himma, and Scott J. Shapiro, 548–87.
Laat, Paul B. de. 2019. "The Disciplinary Power of Predictive Algorithms: A Foucauldian Perspective." *Ethics and Information Technology* 21 (4): 319–29. https://doi.org/10.1007/s10676-019-09509-y.
Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, Mass: Harvard University Press.

———. 2004. *Politics of Nature: How to Bring the Sciences into Democracy*. Cambridge, Mass: Harvard University Press.
Lewis, Andrew. 2010. "User-Driven Discontent." *MetaFilter*. https://www.metafilter.com/95152/Userdriven-discontent.
Liao, Shen-yi, and Bryce Huebner. forthcoming. "Oppressive Things." *Philosophy and Phenomenological Research*, 27.
Lister, Mary. 2020. "All of Facebook's Ad Targeting Options (in One Epic Infographic)." *WordStream* (blog). February 26, 2020. https://www.wordstream.com/blog/ws/2016/06/27/facebook-ad-targeting-options-infographic.
Marr, Bernard. 2016. *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Chichester, West Sussex: Wiley.
Martin, Kirsten. 2019. "Ethical Implications and Accountability of Algorithms." *Journal of Business Ethics* 160 (4): 835–50. https://doi.org/10.1007/s10551-018-3921-3.
Mecacci, Giulio, and Filippo Santoni de Sio. 2019. "Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles." *Ethics and Information Technology*. https://doi.org/10.1007/s10676-019-09519-w.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 205395171667967. https://doi.org/10.1177/2053951716679679.
Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press. https://muse.jhu.edu/book/64995/.
Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24 (4): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.
O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books.
O'Shea, Tom. 2018. "Disability and Domination: Lessons from Republican Political Philosophy: Disability and Domination." *Journal of Applied Philosophy* 35 (1): 133–48. https://doi.org/10.1111/japp.12149.
Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.
Pettit, Philip. 1996. "Freedom as Antipower." *Ethics* 106 (3): 576–604.
———. 2002. *Republicanism : A Theory of Freedom and Government*. 1 online resource (x, 304 pages). vols. Oxford Political Theory. Oxford: Clarendon Press. http://site.ebrary.com/id/10273243.
———. 2012. *On the People's Terms : A Republican Theory and Model of Democracy*. 1 online resource vols. The Seeley Lectures. Cambridge ; Cambridge University Press. https://doi.org/10.1017/CBO9781139017428.
———. 2014. *Just Freedom : A Moral Compass for a Complex World*. First edition. The Norton Global Ethics Series. New York: W.W. Norton & Company. http://bvbr.bib-

bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=027216451&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.
Rini, Regina. 2018. “How to Take Offense: Responding to Microaggression.” *Journal of the American Philosophical Association* 4 (3): 332–51. https://doi.org/10.1017/apa.2018.23.
Rubel, Alan, Clinton Castro, and Adam Pham. 2019. “Agency Laundering and Information Technologies.” *Ethical Theory and Moral Practice* 22 (4): 1017–1041. https://doi.org/10.1007/s10677-019-10030-w.
Sadowski, Jathan, and Frank Pasquale. 2015. “The Spectrum of Control: A Social Theory of the Smart City.” *First Monday* 20 (7): 1–25.
Santoni de Sio, Filippo. 2016. “Ethics and Self-Driving Cars: A White Paper on Responsible Innovation in Automated Driving Systems.” TU Delft. http://resolver.tudelft.nl/uuid:851eb5fb-0271-47df-9ab4-b9edb75b58e1.
Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. “Meaningful Human Control over Autonomous Systems: A Philosophical Account.” *Frontiers Robotics AI* 5 (FEB): 1–14. https://doi.org/10.3389/frobt.2018.00015.
Simons, Jon, Christopher Falzon, Timothy OLeary, and Jana Sawicki. 2013. “Power, Resistance, and Freedom.” In *A Companion to Foucault*. Chicester, West Sussex, UK: John Wiley & Sons Ltd.
Skinner, Quentin. 2004. *Visions of Politics. Volume 2, Renaissance Virtues*. 1 online resource vols. Cambridge ; Cambridge University Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=112400.
———. 2010. “On the Slogans of Republican Political Theory.” *European Journal of Political Theory* 9 (1): 95–102. https://doi.org/10.1177/1474885109349407.
———. 2012. *Liberty before Liberalism*. 1 online resource (156 pages) vols. Canto Classics. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139197175.
Snowden, Edward J. 2019. *Permanent Record*.
Sparrow, Robert, and Mark Howard. 2017. “When Human Beings Are like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport.” *Transportation Research Part C: Emerging Technologies* 80 (July): 206–15. https://doi.org/10.1016/j.trc.2017.04.014.
Vanolo, Alberto. 2014. “Smartmentality: The Smart City as Disciplinary Strategy.” *Urban Studies* 51 (5): 883–898. https://doi.org/10.1177/0042098013494427.
Vincent, Nicole A. 2011. “A Structured Taxonomy of Responsibility Concepts.” In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole A. Vincent, Ibo van de Poel, and Jeroen van den Hoven, 15–35. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_2.
Winner, Langdon. 1977. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, Mass: MIT Press.
———. 1980. “Do Artifacts Have Politics?” *Daedalus* 109 (1): 121–36.
Wood, David Murakami. 2007. “Beyond the Panopticon? Foucault and Surveillance Studies.” In *Space, Knowledge and Power: Foucault and Geography*, 245–263. Aldershot:

Ashgate Publishing.
Yeung, Karen. 2018. “Algorithmic Regulation: A Critical Interrogation.” *Regulation & Governance* 12 (4): 505–23. https://doi.org/10.1111/rego.12158.