Annemarie Sint Jago 0205753

A peek into the discursive construction of the Google Search Algorithm: A critical discourse analysis

Google Search I'm Feeling Lucky

Annemarie Sint Jago 0205753

Google

A peak into the discursive construction of the Google Search Algorithm: A critical discourse analysis

Google Search I'm Feeling Lucky

Master thesis New Media & Digital Culture

Annemarie Sint Jago
Student number: 0205753
a.sintjago@students.uu.nl

3 January 2016

Examinators:
dr. Imar de Vries (1st reader)
dr. Stefan Werning (2nd reader)

Table of Contents

Summary

Algorithms are increasingly interwoven with our daily lives. They are everywhere; yet, they are often invisible. Many people are not aware (of the extent) that algorithms decide what to show and what not to show on their Facebook feeds; or that prices for an overnight stay may depend on which browser is used to look for accommodations. While daily life is saturated with algorithms, just few people have access to its actual codes, enabling them to glimpse into or influence what it does and why. Yet, there seems a consensus that algorithmic opacity might lead to unequal power distribution and abuse of power. This thesis outlines current discussions on search algorithms that address possible implications of algorithmic opacity, such as: bias, discrimination, transparency, (other) concepts that are related with power, and the position of the (cultural) sciences in this field. This framework renders a context for the research of this thesis, which attempts to identify how the Google Search algorithm is discursively constructed by Google and how this discursive construction may reflect power relations between this powerful software giant and its users. Both quantitative and qualitative methods of critical discourse analysis were used to study Google's statements about Google as a company and its search products. It was found that Google's statements were often contradictory: explicit statements were undermined by linguistic structures and specific uses of words that contradicted explicit utterances. Moreover, Google uses words such as 'natural' to make their products and processes more agreeable, while mitigating language is used when negative aspects of Google's search algorithm were mentioned. Lastly, it was found that some topics were completely absent from the Google's discourse, such as bias, discrimination, algorithmic transparency and ethics, which are topics that are deemed important by academics. This thesis closes with remarks of the difficulty of researching algorithms due to its complex networked system and suggestions for further research.

Jennifer was looking for a job. After several attempts to find a job by registering as a job seeker at three employment agencies and days and days of wandering the streets to scan employment agencies' windows in the city she lived in, Jennifer decided to change her tactics and exchange the world of flesh and blood for the one of ones and zeroes: THE INTERNET—half of the job offerings that were taped at the employment agencies' windows referred to websites anyway. So why shouldn't she look for jobs on the Internet? Scanning showcases was so two thousand and two anyway.

In the beginning of her online job-hunt, she decided to type in the URL of a well-known national employment agency, the one with that reliable blue logo. After some unsuccessful browsing sessions on the agency's site, she decided to shift her quest to Google. "www.google.com", she entered into her browser's address bar.

She found herself facing an almost completely white page, with a promising empty box in the middle, hovering over two, what seemed like, buttons: "Google Search" and "I'm feeling lucky", she read. Yes, she was feeling lucky. Who knew what she would find once she had filled the text box with words?!

In the box, which now appeared to her almost magical, a flickering cursor seemed to be inviting her, as if saying "come on then, don't be afraid, I'm not waiting for you forever. Go!" Jennifer took a deep breath, put her fingers on her laptop's keyboard, entered what she was looking for and pressed enter, ignoring the buttons below the text box.

Coincidentally, her brother was also looking for similar jobs and one evening, Jennifer and her brother were sitting at Jennifer's dining table, they found themselves entering exactly *the same keywords on their laptops. And it was just after the moment that they had pressed* ENTER *that it got really weird.*

1 Introduction

Algorithms can rig, manipulate, and discriminate. This is what three recent studies on search algorithms claim (Epstein and Robertson 2015; Epstein 2015; Datta, Tschantz, and Datta 2015; Kay, Matuszek, and Munson 2015). In a study seeking to understand what criteria are used by Google's search algorithm and how those criteria affect types of ads that are displayed in search results, [1] Datta et al. found that male jobseekers were presented different ads than female jobseekers. However, Datta et al. note that the "advertising ecosystem" is complex and hence they cannot determine whether the discrimination in ads presentation is done by Google, specific ad companies, or by other players in this network (Datta et al 2015, 93). Nevertheless, regardless of whether Google's algorithm is responsible for the selection of job ads in the result page, or whether this is due to set criteria by ad companies, men received much higher paying jobs in their search results than women. Some people would call that discrimination in the normative sense.

While we can all imagine that showing high paying jobs only to specific groups may be ethically and morally problematic, a recent study scrutinized a related, seemingly less controversial, phenomenon that can also have serious consequences: Kay et al. found that gender distributions of specific occupations in image search results were unrepresentative of gender distributions in the real world, and that these search results can influence people's perception of gender distribution in reality (the real world) (2015). To be specific: people "believe results are better when they agree with the stereotype" (Kay, Matuszek, and Munson 2015). For instance, participants of Kay et al.'s research regarded a male construction worked as more professional than a sexy dressed female construction worker (Kay, Matuszek, and Munson 2015). Kay et al. add that bias does not only appear in search results, but there can also be bias in the autocomplete suggestions of the search field, thus even before a search query is executed.[2]

---

[1] Specifically for this study, Datta et al. developed a tool, AdFisher, to study the causal relation of gender specific properties and browsing behaviour with ads served by Google. The tool was used to automatically determine the difference between the ads served for men, and the ads served for women, while so far related research has manually looked for markers of discrimination of different groups (Datta, Tschantz, and Datta 2015, 94–95). Basically, machine learning was used to determine signifying features that distinguish two or more groups (102). Hence, by letting the computer look for patterns that determined whether the user was female or male, they found that these patterns were specific websites or ads. In other words: if a user was getting an ad about construction work, the computer could predict that the user was probably male. Using this information, it appeared that males were served ads such as construction work, whereas women didn't get these ads.

[2] Furthermore, taking even a step back from the search of images, Kay et al. claim that there are also "pre-existing biases that affect the images available for image search systems" (Kay, Matuszek, and Munson 2015). Although Kay et al. do not go into this aspect, I think that they mean that bias operates not only in the process of distribution (getting the images to the viewer by using a search engine, for instance), but that the production of images is already biased: for instance, the selection of which images will and will *not* be produced (see: Frosh 2001).

Kay et al.'s findings that results that agree with people's existing ideas (such as stereotypes) strengthens existing ideas is also acknowledged by Eli Pariser. He discusses implications of algorithms showing what they think we want to see, which he calls the "filter bubble" (Pariser 2011). Hannak et al., summarising Pariser, assert that personalization may result in filter bubble effects, that is, "users are only given results that the personalization algorithm thinks they want (while other, potentially important, results remain hidden)" (Hannak et al. 2013). They add that the "Filter Bubble effect is exacerbated by the dual issues that most users do not know that search results are personalized, yet users tend to place blind faith in the quality of search results" (Hannak et al. 2013). Kay et al conclude from a literature study on stereotyping: "The information people access affects their understanding of the world around them and the decisions they make: biased information can affect both how people treat others and how they evaluate their own choices or opportunities" (2015).

Another recent study on algorithms and ethical implication was done by Epstein and Robertson (2015). In "The Search Engine Manipulation Effect (SEME) and Its Possible Outcomes of Elections" Epstein and Robertson found that users were influenced by bias in search results and ranking in search results, even when users knew that results were biased (2015, 4519). Epstein and Robertson concluded that a search algorithm may therefore be a powerful means that can be abused (Epstein 2015). They also address the implications of personalisation by search algorithms. Search engines can access specific groups, thereby making detection and regulation difficult (Epstein and Robertson 2015, 4519).

Studies of algorithms are not limited to merely academics, or law;[3] the increased dependency on algorithms, its spreading use, and its increasing density in our daily life raises questions and concerns in many areas. Yet, while algorithms are part of daily life, only few people have access to its workings and can get insight in how they work, what kind of criteria are used, etc. Most users do not even know when algorithms interfere with their activities (Hamilton et al. 2014). It is not surprising therefore that the last years, a growing body of research addresses issues of ethics, bias, transparency, etc.

Claims and discussion such as these, and other discussions concerning algorithmic bias, abuse of power, contrast the early promises of the Internet: namely, the promise of Internet offering the human race total equality. However, "online discourse is woven of stereotypical cultural narratives that reinstall precisely those positions [of race, gender, and class]" (Daniel Punday quoted in Baker and Potts 2013, 187). "A decade ago, the Internet was frequently viewed through a utopian lens, with scholars predicting that this increased ability to share, access, and produce content would reduce barriers to information access" (Granka 2010, 364). The Internet seems thereby not different from other technologies; many systems of technology have at the time been described as "democratizing, liberating forces" (Winner 1980, 121): "[s]carcely a new invention comes along that someone does not proclaim it the salvation of a free society" (Winner 1980, 122).

[3] A few years ago, Google was accused because they were suspected of favouring their own services in their shopping results (Wouter Van Noort 2015).

Hinman describes search engines as "principal gatekeepers of knowledge" (Hinman 2005, 25). Without search engines we would not be able to get the information we are looking for on the Internet. Hinman describes how the initial search engines used objective criteria to deliver the search results to the user: these search results were based on the popularity of sites (which was based on the number of page views), incoming links etc. (Hinman 2005, 22). Later, search engines took into account the needs of the user: "what the user wants becomes an integral part of the formula, as does the set of search terms most commonly used to express what the user wants" (Hinman 2005, 22). One of the dangers of search engines being the gatekeepers of knowledge, is the fact that the companies that own these engines are not neutral: they have their own agenda and might be influenced by powerful parties (Hinman 2005, 22–23), which is also stated by Epstein who fears that this possibility and manner of exerting power may even be a threat for democracy (Epstein 2015). Moreover, the search process itself is opaque, so the public does not know why some results are included and why some are excluded (Hinman 2005, 22) (or do not even know which sites are excluded):

> They [search engines] are like windows onto the web—and, like windows, tend to be largely unnoticed because our gaze focuses on what is visible through them. With windows, however, it is easy to detect when they are cloudy or distorted. With search engines, however, it is much more difficult to tell when they are providing distorted or incomplete pictures. (Hinman 2005, 21)

While Hinman describes the algorithm as a window, the algorithm may also be a *mirror*. Marcus Moretti claims: "In some cases they [Google Image Search Results] can be seen to reflect the stereotypes, preconceptions, attitudes and ideals of the American Internet-browsing public" (Moretti 2014).

Whatever the discussions may be, if we shift our attention to the algorithm itself, we see that the algorithm itself is inherently paradoxical, which itself is already sufficient material for discussions. "Autonomous decision-making is the crux of algorithmic power", Diakopoulos notes (Diakopoulos 2014, 3), that is, decision-making without human intervention. While the algorithm makes decision on its own, it performs these actions unseen. While good interaction designs usually involves visibility (the user does not need to be aware of every process), in the case of algorithms, this invisibility is a problem: "man [scholars] see opacity in technologies as a call for inquiry into what processes of debate and concern have been arrested or settled behind the opaque surface of 'black boxes'" (Hamilton et al. 2014).

The invisible and autonomous aspects of the decision making process are exactly what is at the core of current discussions on the algorithm. Most scholars propose possible steps or actions for how to proceed—in which the disclosure of algorithms is foremost proposed. For instance, Introna and Nussbaum demand that "underlying rules (or algorithms) governing indexing, searching and prioritizing" should be fully and truthfully disclosed "in a way that is meaningful to the majority of Web users" (Introna and Nussbaum 1999, 34). Chapman and Rotenberg argue for "the Development of standardized methods for information finding" (quoted in Introna and Nussbaum 1999, 29). Yet, while disclosure of algorithms is foremost proposed, scholars also acknowledge that disclosure is not realistic, since that would open the way for spammers and hackers (e.g. Sandvig et al. 2014, 9; Introna and Nissenbaum 1999, 34; Granka 2010, 366). Sandvig et al., for instance, argue that public disclosure of algorithms "might be likely to produce serious negative consequences" (2014, 9). They add:

> On many platforms the algorithm designers constantly operate a game of cat-and-mouse with those who would abuse or "game" their algorithm. These adversaries may themselves be criminals (such as spammers or hackers) and aiding them could conceivably be a greater harm than detecting unfair discrimination in the platform itself. (Sandvig et al. 2014, 9)

Scholars seem to depart from the assumption that algorithms need to be opaque to a certain extent to prevent abuse. Yet, is this need really true, or is it just an example of powerful framing of algorithms? Why do scholars accept not giving full disclosure that easily?

This thesis seeks to understand how Google uses discourse to construct, maintain and reinforce assumptions and presuppositions concerning the politics of the algorithm—such as the assumption that algorithms must be opaque to avoid abuse by spammers. In other words: how is the Google Search algorithm discursively constructed in relation to user choice, bias, transparency, and what are implications of this discursive construction for the conservation of power relations between Google and its users?

To answer this question, I have used theory and methodology borrowed from Critical Discourse Analysis (CDA). A corpus of Google texts was composed, on which I performed quantitative and qualitative analysis. The quantitative analysis entailed word frequency queries, while the qualitative analysis was done by close reading as a part of Norman Fairclough's approach of CDA. In chapter 2 I will describe the methodology more extensively. It was decided to focus on the Google Search algorithm specific, since it is the most used search engine by people worldwide. It is thereby one of the most used mediators between user and knowledge. This means that possible discursive constructions of this algorithm have huge user impact and reach.

Firstly, in chapter 1 ("Defining the search algorithm in context of current debates") I will elaborate on the algorithm, what it is and why studying 'the algorithm' is complex. Is there such a thing as 'the algorithm', for instance? I will also outline and extend on previous research on search algorithms to understand how my research is positioned in current discussions on the algorithm and to understand the context that is needed to later interpret the findings from the qualitative analysis. In the subsequent chapter ("Methodology and methods") the CDA methodology, including its theoretical underpinnings will be discussed, which includes concepts such as power, hegemony, discourse, and language (chapter 2.1 "Theoretical underpinnings"). In this first section also the used methodology, Critical Discourse Analysis, will be discussed: its history, its critique, and the specific approach of CDA that was used in this study, namely Norman Fairclough's. The second section (chapter 2.2 "Methods of data collection and analysis") describes the methods of data collection and data analysis of the quantitative and qualitative methods. In the subsequent chapter (chapter

3 "Analysis/ interpretation") the actual results of my research are found. Consequently, I will reflect on my research and its limitations (chapter 4: "Reflection and limitations") and close with chapter 5 ("Conclusion: from research findings to future research") in which I will discuss how to proceed from here and give suggestions for further research. Defining the search algorithm in context of current debates

1.1 What is an algorithm? What is considered (part of) an algorithm?

Firstly, to clarify what I exactly am going to research, *and* to grasp the complexity of what elements / actors take part in the constitution of 'the algorithm', I will firstly consider what a search algorithm *is*. In other words: if we talk about a search algorithm, what (parts) do we refer to? What is it? It must be noted that when I refer to *the* algorithm, which is singular, I mean the *concept* of the algorithm, not a singular algorithm, since most algorithms are in fact not one algorithm but consist of many algorithms.[4]

Technically speaking, an algorithm is a series of (mathematical) steps that are finite and "lead to the transformation of some data" (Ince 2009). It is also argued that algorithms do not have to be computational: Rubinstein and Sluis argue that an algorithm is essentially a subsequence of steps and that in this sense analogue photography, for instance, was already algorithmic: the photographer performed structured steps in the darkroom to produce a certain outcome (Rubinstein and Sluis 2013, 28). Algorithms can also consist of series of instructions that have to be followed by a physician to decide on a treatment strategy (such as diagnostic and therapeutic algorithms) ("Algorithm" 2010). In this thesis, I will focus on *computational* algorithms, specifically the Google Search Algorithm. An example of a computational algorithm is a formula that takes as input X, divides X with 2, and displays its result. A simplistic diagram would look as depicted in Figure 1. The algorithm itself would be the formula that executes the steps: take input A, divide it by 2, print result on screen.

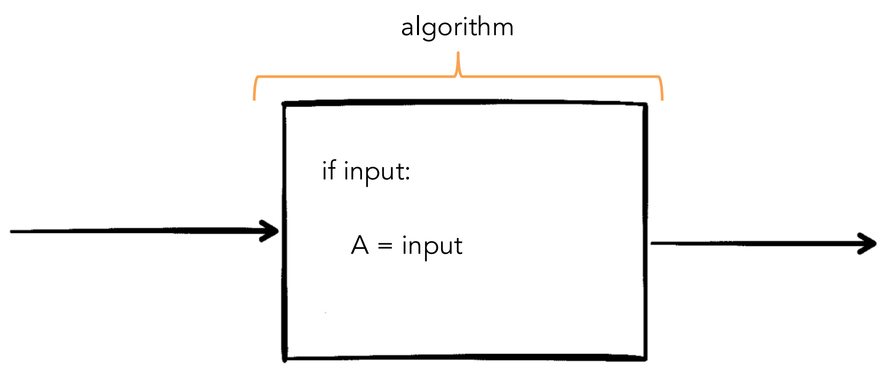

Figure 1: simplistic diagram of a computational algorithm

[4] The Google Search Algorithm, for instance, consists of many algorithms, including its famous PageRank algorithm ("Algorithms" 2015).

Stating that the algorithm is merely a code that can be executed may be shortsighted: without input, the algorithm would not be able to do anything and in a way, the algorithm is only functional when it has an input *and* some output (whether or not the output can be seen). The output can be a printed output on a screen, or it can be something that is stored as a variable. In this sense, the algorithm can be understood as the totality of input, algorithmic code, *and* output, as is simplistically represented in Figure 2.

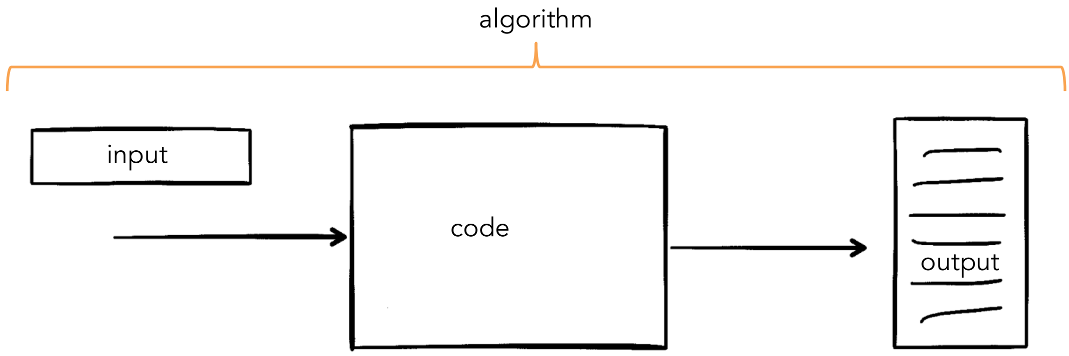

Figure 2: The algorithm is not merely a mathematical procedure; it is inseparable of its input and output.

In case of the Google Search Algorithm, the algorithm needs the data from the Google index[5], such that if a user searched for "history Netherlands", the (simplistic) algorithm would use as input the user's search query *and* the Google index. In the simplified diagram above, another input would thus need to be added, namely the Google index that serves as another input.

Yet, what can be used as input is neither limited nor static. As a matter of fact, the output that the search algorithm generates (a list of links to pages) is used as new input for the algorithmic procedures: if a users clicks on a link to a webpage, information about this user action is put back into the algorithm, so that the algorithm can 'learn'. Moreover, the algorithm also uses other user data, such as the user's browsing history and geographical information ("Algorithms" 2015; Baker and Potts 2013; Diakopoulos 2014; Dwork et al. 2011; Hannak et al. 2013). In fact, these user data do not merely influence the algorithm, in the sense that there is a unilateral connection, the opposite is also the case: a change in the algorithm could also mean that different parts of the input data are used, or that the algorithm makes different choices resulting in a different output. A simplified diagram would then look like Figure 3:

[5] Basically, the Google index is, as its name suggests, an index of words and their locations of pages on the web ("Crawling & Indexing" 2015).

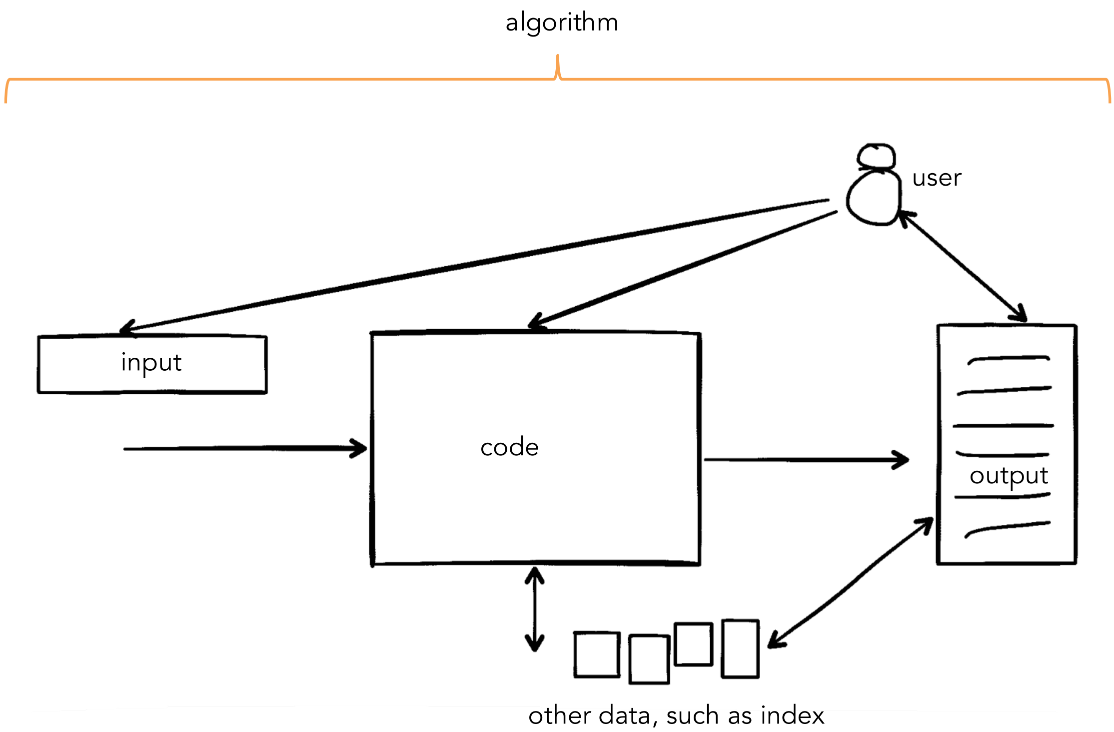

Figure 3 The algorithm as a whole network of codes, inputs, outputs and engaged users. Note that developers, or other possible actors, are not even included.

Note that the output affects the user, but the user also engages with the output, which has a possible effect on the algorithm, and also on the rankings (below). In this simplified scenario, the algorithm includes input, output, as well as user behaviour and user information.

Algorithm and (external) data are, thus, inextricably linked. As Sandvig et al. argue: the algorithm needs data, otherwise disclosure of algorithms would not mean anything, since algorithms may only produce specific output in the context of a particular data set (Sandvig et al. 2014, 10). If you would want to study an algorithm, you therefore also need to know what data is used to see how the algorithm works. Naturally, the real Google Search algorithm is much more complex than outlined in Figure 3 since it is based on more than 200 'signals' and other algorithms such as Google's PageRank ("Algorithms" 2015).

1.2 Current discussions on search algorithms

As can be seen from the simplified diagram, many connections are established between different actors, which complicates studying 'the search algorithm', or even the Google Search Algorithm, as if it were a static object. The algorithm, in this sense, is rather a networked object. Research has focused on various parts of this 'algorithmic network', such as studies that focus on output and user. The following overview of previous studies on search algorithms, bias and discrimination shows just how complex the topic of the search algorithm is and the many perspectives to it, which are all in a way connected.

1.2.1 Bias, stereotyping, reinforcement of stereotyping

Epstein and Robertson (2015) zoom in on the relation between output and user. Continuing on earlier research that found that ranking of search results influence users, Epstein and Robertson have studied whether users of a mock-up search engine were influenced by biased search results on political candidates (Epstein and Robertson 2015). They conclude that a type of bandwagon effect is created: search rankings are affected by voter preferences, while voter preferences are affected by search rankings, resulting in magnification of minor differences (Epstein and Robertson 2015, 4519). Magnification of bias, inequalities, differences, or stereotypes was also found in other studies (Kay, Matuszek, and Munson 2015; Pariser 2011).

Kay et al. (2015) also studied the relation between search results—in this case: image search results—and user perception of the real world. They found that gender distributions of specific occupations were unrepresentative of the gender distributions in the real world *and* that search results can influence people's perception of gender distribution in reality (Kay, Matuszek, and Munson 2015). Participants in Kay et al.'s study considered stereotypes of professions more professional and appropriate than images that were not stereotypical (2015). Kay et al. state that unrepresentative gender distribution in search results "risks reinforcing or even increasing perceptions of actual gender segregation in careers" (2015). However, their study does not (and probably does not seek to) explain whether unrepresentative search results are a consequence of an unrepresentative collection of images that the search engine has at its base, or whether it is the algorithm that is producing unrepresentative results. As discussed in the introduction, Kay et al. posit the probability of pre-existing biases that affect search results, such as representativeness of the available data that is searched, and search suggestions of the autocomplete function.

Bias in auto-completion has been studied by Baker and Potts who have focussed on discriminatory effects of the auto-complete search algorithm of Google (2013). Baker and Potts state, extending on (2011) and Pariser, that learning is about being confronted with new things. The filter of Google hinders the encounter with new things: "people can [become] enclosed in a loop, whereby they are only directed to sites which they have previously shown interest in" (Baker and Potts 2013, 188). The filter bubble "tends to dramatically amplify confirmation bias" (Pariser 2011). Baker and Potts found that some auto-complete suggestions were considered inappropriate/racial/negative, have unintended consequences such as reinforcement of negative stereotypes (201), and that Google has no function to flag inappropriate auto-complete suggestions (200). Baker and Pots hold that the auto-complete suggestions "offer a window into the collective Internet consciousness" (201).[6] How to cope with this is difficult. Should particular racial questions be removed, Baker and Potts ask? They bring forth possible counterarguments of censoring the Internet and the

[6] While Baker and Potts think that the majority of Internet users are not asking negative stereotypical questions, the auto-complete suggestions reveal that enough people are asking particular questions for the auto-complete suggestions to offer these questions (Baker and Potts 2013, 201).

problem whom should decide about appropriateness (201-201). One solution could be that Google should remove suggestions that are consistently flagged.

Epstein and Robertson conclude from earlier research that: “most people are relatively powerless when trying to resist sources of influence they cannot see” and “when people are unaware they are being manipulated, they tend to believe they have adopted their new thinking voluntarily” (Epstein and Robertson 2015, 4520). Even when users are aware of bias, they are influenced by rankings (Epstein and Robertson 2015, 4519).

To make matters of bias and discrimination even more complex, there is also the advertising aspect that is involved in search: search engines are not merely limited to searching webpages, but advertisements and advertised links are also part of the algorithm; as already discussed in the introduction, Datta et al. found that men were shown high paying jobs, while only few women were shown high paying jobs (Datta, Tschantz, and Datta 2015).

1.2.2 Personalisation

Other research has attempted to learn more about the algorithmic code of the GSA itself and the data that is used by studying the effect of various user data on the algorithm’s output. Hannak et al. have sought to find out “what user features influence Google’s search personalization algorithms” and “to what extent . . . search personalization actually affect[s] search results” (Hannak et al. 2013). They introduce “a methodology for measuring personalization on Web search engines” (Hannak et al. 2013) since personalisation may lead to filter bubble effects: the user is served only information that the algorithm thinks is relevant for the user, while other information is left out. Users may not even be aware of (the extent of) personalisation and blindly trust the search results. Hannak et al. see the “highest personalization for queries related to political issues, news, and local businesses” (Hannak et al. 2013). Filtered news results, for example, especially when the user is unaware of the filtering, may lead to a distorted perception of events. Consider a hypothetical yet realistic example of Susan, who is interested in the human body and health and therefore reads many health blogs. When Susan searches for information on political parties, the algorithm may decide to put the parties that have strong stances on public health on the top of the search results page. Yet, Susan may find public health irrelevant when it comes to politics. She might merely be interested in health with regard to how to live healthy, while in politics she finds national economic stability crucial. Nonetheless, the algorithm has already ranked political parties for her, based on a different feature, which might influence how Susan will approach these parties.

Other studies have looked at user data and user browsing history in relation to personalized search results. In one of these studies, it was found that the output of the algorithm is different for 16% of the (fictional) users; however, the study could not attribute this to the search algorithm itself, or other entities, such as advertiser’s settings (Datta, Tschantz, and Datta 2015). Drawing these external parties in the diagram (adding them to Figure 3) would make the network even more complex, which shows how difficult it is to study algorithms, since algorithms are so fluid and dependent on various factors.

Additionally, exactly because specific groups or users are targeted it is difficult to detect and regulate personalisation (Epstein and Robertson 2015, 4519). Epstein and Robertson also note that while campaign influence is usually explicit, search engine manipulations are hard to detect (Epstein and Robertson 2015, 4518).

1.2.3 Benchmarks / standards

Where does personalisation ends and discrimination starts? There is no clear standard for this, although there are cases in the US that have been considered as discrimination (Kirchner 2015). Also Google itself has apparently set a line in what is considered acceptable and inacceptable: for instance, there are no auto-complete suggestions for queries that are about pornography or illegal downloading (2015), and some type of images, such as a "racist photoshopped image of Michelle Obama", are not included in image search (Gillespie 2014, 180).

While many researchers claim that algorithms, or its results, are biased or unfair, most scholars, such as Epstein and Robertson[7], neglect the fact that no benchmarks exist, which makes it problematic to consider something as unfair or biased. Other scholars do acknowledge the problem of a lack of standards of what is considered fair and unfair, neutral, bias, discrimination and so forth. Jeremy Kun, for instance, raises the question: "what does it mean for an algorithm to discriminate" and concludes that there is no benchmark for what is considered discrimination and what not (Kun 2015).

Dwork et al. (2011) have attempted to create a tool and algorithms "that guarantee fairness" (Miller 2015). However, in an interview with NY Times' Caroline Miller, Dwork stresses that she approaches fairness from a *mathematical* point of view: "Fairness means that similar people are treated similarly . . . It would require serious thought about who should be treated similarly to whom," she adds (Miller 2015), referring to mathematical classification schemes. She argues that ethicists should determine "whose responsibility it is" to "ensure that algorithms or software are not discriminatory" (Miller 2015).

---

[7] Consider the following hypothetical example in relation with Epstein and Robertson's research: candidate A and candidate B are participating in elections. The PR-person of B has accomplished that lots of media write about candidate B. This would influence search results, because in a specific period, more media have discussed person B than person A. How should a search engine behave? Should it list more results about person B (because recently more news has been about B?), or should it divide the amount of search websites equally, since there are two candidates? And if one of the candidates has been discussed in the news because of a personal incident that has affected many people's attitude towards him or her? Should a search engine distinguish between relevance, or not? Is a search engine biased if it does show results about personal matters, or should it stick to news about politics? And if Americans tend to be more influenced by personal contextual information about candidates (think about the Monica Lewinsky affaire), but Dutch people restrict themselves to the political functioning of a person. How should a search engine decide on relevancy?

1.2.4 Meta-perspective on algorithms

Other studies have stepped outside the functionality of algorithms or its networks, such as Introna and Nissenbaum's article (1999), which focuses on the place/space in which the search algorithm operates. They claim that this place, namely the Web, is a public space (Introna and Nissenbaum 1999). Biased search results can undermine the web as a public space. They rather not see the Web as a free marketplace. Building forth on Elizabeth Anderson, Introna and Nissenbaum state that there are "goods that should not be left entirely (if at all) to the marketplace" (Introna and Nissenbaum 1999, 22); "there are certain goods—ones that Anderson calls 'political goods' and includes among the schools and public spaces—that must be distributed not in accordance with market norms but 'in accordance with public principles'" (Introna and Nussbaum 1999, 23). Introna and Nissenbaum's notion is interesting, since the nature of the Web has increasingly become more hybrid: since the publication of their article the share of commercial content has grown enormously. A possible problem, I would add, is that Internet is at once a public space and a space for advertising. While formerly these were separated in libraries, schools and public spaces, now public and commercial content are sharing the same spaces. Search result pages deliver both public and commercial information in the same place, as if they are the same type of information. This makes dealing with information more complex than when public and commercial information are offered in separated places.

1.2.5 Detection, regulation of biased algorithms

Sandvig et al.'s concern is not whether algorithmic providers are good or evil, but "what mechanisms we have available to determine what they are doing at all" (Sandvig et al. 2014, 17). They propose five algorithm audits: "a starting point for a future conversation about how researchers might look inside the black box of the algorithm to pursue knowledge about pressing public problems" (Sandvig et al. 2014, 8). They have concerns, however, that audits may violate the "basic principles for the ethical conduct of science" (Sandvig et al. 2014, 7), the scrutinized Internet platform's terms of service, other academic codes of conduct, or even the law, in particular the Computer Fraud and Abuse Act (CFAA). They also note that the social problems that they have addressed in their paper (they illustrated some cases of discrimination by algorithms) "are often instances of those problems we have been addressing for a long time" (Sandvig et al. 2014, 17). Surprisingly, the academic community and the law did tolerate traditional *offline* audits that were conducted to reveal discriminatory phenomena or patterns, because they served the greater good (Sandvig et al 2014, 17). Nowadays, while the proposed algorithmic audits basically reveal similar instances of discrimination, these algorithmic audits are prohibited by law and by academic codes of conduct (Sandvig et al 2014, 17). Therefore, Sandvig et al. propose not for transparency or monitoring misbehaviour on Internet, but they rather argue for "regulation towards audibility" (Sandvig et al 2014, 17). This implies reforming the CFAA, scholarly guidelines, and a third-party role for government and the public (Sandvig et al 2014, 17-18). Lastly, Sandvig et al. note that it is difficult to judge discrimination by algorithms while it is unclear how society wants algorithms to behave (Sandvig et al 2014, 18). Also, algorithm audits suffer from the same lack of standards as judging algorithms as biased or unfair: there are no answers to questions such as "What is the minimum amount of data that would be required to detect a significant bias in an important algorithm?" (Sandvig et al 2014, 18). The key question remains: "how do we as society want these algorithms to behave" (Sandvig et al 2014, 18)?

1.2.6 Articulation of the algorithm

Tarleton Gillespie is one of the few scholars who consider the algorithm from a discourse perspective. His notion on articulations touches upon the discursive articulations that co-constitute the algorithm that I address in this thesis:

> The articulations offered by the algorithm provider alongside their tool are meant to . . . *define* their tool within the practices of users, *to bestow the tool with a legitimacy* that then carries to the information provided and, by proxy, the provider. The careful articulation of an algorithm as impartial (even when that characterization is more obfuscation than explanation) certifies it as a reliable sociotechnical actor, lends its results relevance and credibility, and maintains the provider's apparent neutrality in the face of the millions of evaluations it makes. This articulation of the algorithm is just as crucial to its social life as its material design and its economic obligations. (Gillespie 2014, 179, my emphasis)

According to Gillespie, articulations "bestow the tool with . . . legitimacy" and can render it "impartial" (Gillespie 2014, 179): this means that an algorithm is shaped by its surrounding articulations.

> This articulation happens first in the presentation of the tool, in its deployment within a broader information service. Calling them "results" or "best" or "top stories" or "trends" speaks not only to what the algorithm is actually measuring, but to what it should be understood as measuring. (Gillespie 2014, 180)

In this claim, Gillespie touches upon something important: the presentation of the tool, not as something as *doing* particular things, but as how we *perceive* it. To illustrate this with an analogy: consider the terms *freedom fighters* and *terrorists*. This is an example of words that people generally know are ideologically charged; however, it is more difficult to see the underlying ideology of words if you're in the middle of it.

1.2.7 How to proceed from here

The topic of the algorithm touches upon many discussions and contemporary issues. What most studies share—regardless of what part of the algorithmic network, as I would call it, they studied—are the following points:

- The lack of benchmarks. What is fair? When is something considered biased? When are search results relevant, and when not? What is neutrality?
- The need for transparency of algorithms, its data, and used criteria for the selection of data. Yet, while scholars find it troubling that algorithms are opaque, they acknowledge that making it transparent may be worse than not giving disclosure of the Google Search Algorithm, since disclosure would probably result in abuse, spam, etc.

- A need for regulations concerning auditing or regulations of algorithms

Two aspects stand out in current discussions. Firstly, most studies primarily focus on functional aspects of the algorithm or its context (Introna and Nussbaum 1999), or influence on the user (Epstein 2015). Yet, as is the case with many things in life, an object/actor/agent is not only what it does or in which context it is situated, but it is also how it is textually (linguistically, visually) *represented*. Our understanding, acceptance, or lack of acceptance of the algorithm is partly influenced by how the algorithm is positioned or shaped by its creator, users, or other actors. As Fairclough puts it:

> We cannot take the role of semiosis in social practices for granted; it has to be established through analysis . . . Semiosis in the representation and self-representation of social practices constitutes discourses. Discourses are diverse representations of social life which are inherently positioned – different social actors 'see' and represent social life in different ways, different discourses. (Fairclough paraphrased in Wodak 2001, 123)

How aspects of life are represented, is thus dependent on the discourse in which it is represented.

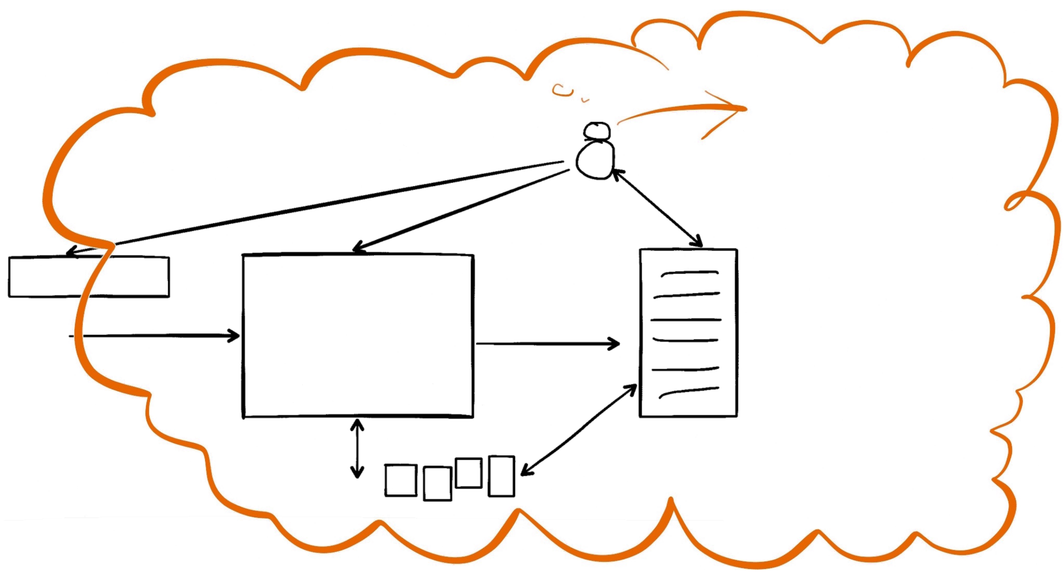

Figure 4: The discourse on algorithms—whether it is focused on the code, its use, or it implications—defines the meaning of the algorithm. The user is imbedded *in* this discourse, which makes it difficult to recognize discursive strategies.

This brings me to my second point: many researchers state that some solutions such as disclosure of the algorithm is not feasible, based on 'arguments' such as possible abuse of disclosed algorithms by hackers (Sandvig et al. 2014, 9), or that disclosure of algorithms will result in search engines returning similar search results (Granka 2010, 366). In this thesis, I will seek how the algorithm is discursively constructed and if and how that explains why scholars seem to think that full disclosure is not feasible. These beliefs may have been constructed or reinforced by a powerful discourse. The lack of algorithmic transparency is in this sense accepted, however reluctantly. In the next chapter, I will account for the methodological approach that has been taken in this study.

2 Methodology and methods

Critical discourse analysis is productive to research contexts in which language is used by powerful agents to construct or hide certain ideas or topics. My thesis aims to understand exactly this phenomenon: how the Google Search algorithm is *discursively constructed* and the implications of this discursive construction *for the conservation of power relations* between Google and its users. Therefore, CDA was chosen as a methodology.

CDA is both a theory *and* a methodology (Wodak 2001, 121). For reasons of structure this chapter has been divided into two sections. The first covers CDA's theoretical underpinnings: it's history, most important concepts, the pros and cons to CDA, and the specific approach that was used in this study. The second section covers the actual methods that were used for data collection and analysis.

2.1 Theoretical underpinnings

How we perceive the world depends partly on the glasses we look through: post-structuralists hold that language constructs how the world is seen and understood (Lister et al. 2009, 68). Consider, for instance, water: "'Water' is not the same theoretical object in chemistry as it is in hydraulics", yet the substance is the same (Viktor Burgin's example in Lister et al. 2009, 68). Similarly, 'freedom fighter' and 'terrorist' may denote the same person, while this person is perceived differently, based on the word that is used to represent him or her. Discourses are "elaborate systems of language" (Lister et al. 2009, 68). The term 'discourse' has been developed by Michel Foucault (Wurth and Rigney 2009, 93). He defines discourse as "practices which systematically form the objects of which they speak" (Foucault 1972, 54). Discourses, thus, also "shape the way we think about new media" (Lister et al. 2009, 77), such as algorithms. Critical Discourse Analysis (CDA) is a theory and a method to study (such) discourses.

CDA, or rather Critical Linguistics (CL), as it became to be known, emerged in the 1970s. It "regards language as a social practice" (Wodak 2001, 1) and it recognized that language played a role in "structuring power relations in society" (Wodak 2001, 5). CL/CDA is consequently particularly interested in "the relation between language and power", especially when there are relationships of inequality, dominance, power and control manifested in language (Wodak 2001, 2). In first instance, CL's focus was mainly on the formal aspects of language; the attention to power and social hierarchy was limited (Wodak 2001, 5). By the 1990s a distinct approach to CL had developed, first by Norman Fairclough (Baker and Ellece 2011, 26), what came to be known as CDA, which was radically different from traditional linguistics (Wodak 2001, 5). This paradigm sees, for instance, language as a social phenomenon and "not only individuals, but also institutions and social groupings have specific meanings and values, that [sic] are expressed in language in systematic ways" (Wodak 2001, 6).

Critical Discourse Analysis does not merely focus on language[8] itself, but it also inquires its context of production, history, how meaning is created in the interaction with texts (Wodak 2001, 2–3). Therefore, "the concept of power, the concept of history, and the concept of ideology . . . figure indispensably in all CDA" (Wodak 2001, 2–3). In many discourses, the powerful determine what is considered normal, while the less powerful echo the powerful (Wurth and Rigney 2009, 93). Proponents of CDA claim that "dominant structures stabilize conventions and naturalize them"; "discourse is structured by dominance" and dominance structures are legitimated by ideologies of powerful groups" (Wodak 2001, 3).[9] CDA offers a method to analyse these processes and possibilities of resistance (Wodak 2001, 3).

It is exactly at the points of the politics of language that the critical in Critical Discourse Analysis comes in to practices. What is critical in CDA is understood differently by scholars (Wodak 2001, 9): some scholars see 'critical' in the sense of adhering to the Frankfurt school or Marx's notions (Wodak 2001, 9). According to Ruth Wodak: "Basically, 'critical' is to be understood as having distance to the data, embedding the data in the social, taking a political stance explicitly, and a focus on self-reflection as scholars doing research" (Wodak 2001, 9). I will come to the matter of reflection later when I discuss critique towards CDA. Nonetheless, it is the "political stance" that Wodak mentions that is by most CDA scholars considered the element that distinguishes Discourse Analysis from Critical Discourse Analysis. Ruth Breeze concludes the following:

> the general consensus is that Critical Discourse Analysis contains two essential elements: A more or less political concern with the workings of ideology and power in society; and a specific interest in the way language contributes to, perpetuates and reveals these workings. (Breeze 2011, 495)

Discourse is connected to language and ideology. Language "is one way that ideologies are constructed, maintained and challenged" (Baker and Ellece 2011, 57). Ideology "can be thought of as the set of ideas belief and aims that a person or group holds" (Baker and Ellece 2011, 57). However, ideologies cannot be simply "read off" from texts, because "meanings are produced through *interpretations* of texts" (Fairclough 1992, 88–89, my emphasis). An utterance can have many interpretations, depending on its context: "'Can you carry the suitcase' could be a question, a request or order, a suggestion, a complaint, and so on" Fairclough illustrates (Fairclough 1992, 82). Consequently, language should not be analysed in isolation, but as it is imbedded in its context and part of interaction (Fairclough 1989, 26), as I will go into further in the methodology section of this chapter.

---

[8] Language is not only considered to be speech, or written text, but it can also be visual or spatial (Fairclough 2001b, 22)). What is considered text, or language, however, is dependent on the approach to CDA (Baker and Ellece 2011, 150). Some discourse analysts consider "text as *anything* that generates meaning, while others see text as only linguistically (Baker and Ellece 2011, 150).

[9] It must be noted, however, that power "does not derive from language": "language can be used to challenge power, to subvert it, to alter distributions of power in the short and long term" (Wodak 2001, 11).

2.1.1 Critique on CDA

Two aspects of CDA that have been subject of critique are CDA's methodology, which allows for multiple, flexible approaches, and the political stance that is taken by the researcher towards the topic that is researched. Henry Widdowson has criticized CDA for its vague definition an its 'biased' approach: that is, he claims that no-one really knows what discourse is, and there is no clear demarcation between text and discourse (Meyer 2001, 17). He also claims that CDA is biased: "it is prejudiced on the basis of some ideological commitment, and then it selects for analysis such texts as will support the preferred interpretation" (Meyer 2001, 17). In other words, the researcher may select the text that will substantiate the hypothesis.

Generally, CDA's general response to critique on CDA such as Widdowson's is that, firstly, the idea that a researcher can be neutral is a fallacy. CDA advocates reflexivity, "so the researcher reflects on his or her own position and how it develops as the research progresses" (Baker and Ellece 2011, 27). In this way, I understand that reflexivity creates a meta-context of the research itself, allowing the research to be assessed by other researchers. Secondly, CDA encourages researchers to include triangulation in their research (Baker and Ellece 2011, 27). Triangulation is the use of multiple approaches, e.g.: using different methods for the collection of data or combining qualitative and quantitative analyses. Triangulation "facilitates validity checks of hypotheses, it anchors findings in more robust interpretations and explanations, and it allows the researcher to respond flexibly to unforeseen problems and aspects of the research" (Layder in Baker and Ellece 2011, 154). In this study, it was attempted to incorporate both reflexivity (included in) and triangulation (see section 2.2 "Methods of data collection and analysis" and chapter 5, "Conclusion") to avoid (confirmation) bias as much as possible.

While I agree with CDA's general response to its critique of biasedness, it seems that in Widdowson's statement (which is exemplary of the general critique on CDA) one important things is ignored: scholarship. I would argue that researchers have and take a certain responsibility towards society. Researchers have been educated to consider multiple perspectives and interpretations in their research. Stating that researchers would substantiate their hypothesis by biased selection of their data is undermining scholarship itself and is not limited to specific methods or areas of research.

2.1.2 Fairclough's approach to CDA

Using CDA to study a particular subject is not like following specific order of predefined steps that will guide you through the research project. There is no single approach to CDA, but there are many (Meyer 2001, 14), even within specific forms of CDA (Baker and Ellece 2011, 26). All these forms of CDA are "strongly based in theory" (Meyer 2001, 17). There are approaches that come from micro-sociologies (Ron Scollon), from "theories on society and power in Michel Foucault's tradition" (Norman Fairclough, Ruth Wodak, Siegfried Jäger), grammar, and social cognition (Teun van Dijk).[10]

For this study, I have used Fairclough's approach as a basis for my methodology. As previously indicated, Fairclough's approach is concerned with theories on power and his approach also includes written text (whereas other CDA approaches are rather concerned with spoken text). His approach to CDA is based on upon Bhaskar's explanatory critique framework (Fairclough 2001a, 124–125). This framework consists of five stages:

Stage 1: Focus upon a social problem which has a semiotic aspect.
Stage 2: Identify obstacles to it being tackled, through analysis of
  a. the network of practices it is located within
  b. the relationship of semiosis to other elements within the particular practice(s) concerned
  c. the discourse (the semiosis itself)
    • structural analysis: the order of discourse
    • interactional analysis
    • interdiscursive analysis
    • linguistic and semiotic analysis
Stage 3: Consider whether the social order (network of practices) in a sense 'needs' the problem.
Stage 4: Identify possible ways past the obstacles
Stage 5: Reflect critically on the analysis (1-4) (Fairclough 2001a, 124–125)

The first stage focuses on defining the problem. What is the problem and for whom is it a problem (Fairclough 2001a, 125)? In this stage of the analysis the researcher needs "to go outside of the text, using academic and non-academic sources to get a sense of its [in this case: the algorithms] social context" (Fairclough 2001a, 129).[11] The findings of this part of the analysis have already been described in the introduction and chapter **Error! Reference ource not found.**, "**Error! Reference source not found.**", since this analysis was also part of a (preliminary) literature review. To recapitulate: algorithms are part of our lives, even when we are unaware of them; the problem is that we do not know *if*, *when* and *how* algorithms play a role in our (daily) life, since algorithms are opaque and powerful companies try to keep them opaque.[12] This lack of transparency is a problem for governments, law enforcers, scholars and the public: governments cannot monitor algorithms, law enforcers idem, scholars do not have access to algorithms to study them—in some cases they are even breaking the law when attempting to study algorithms (Sandvig et al. 2014, 12)—and the public does not know what information about them is collected and used by algorithms and whom has access to that information. This is a problem for all of the above, since abuse of algorithms can go undetected, or unintended problems may arise, such as algorithmic bias. While many studies address this problem and many scholars give suggestions on how to

[10] See Meyer (2001, 19–23) for concise overview of these approaches.

[11] Context "helps in the interpretative process of linguistic phenomena as well as providing explanations" (Baker and Ellece 2011, 21).

[12] With few exceptions, such as Netflix.

proceed, it seems that the lack of disclosure of algorithm is in a way accepted, though reluctantly: transparency leads to abuse, seems to be the core of many arguments (e.g. Sandvig et al. 2014, 9; Introna and Nissenbaum 1999, 34; Granka 2010, 366). It is as if this idea is unquestionable, a fact. The problem is that possible future ways to deal with algorithmic transparency are beforehand (partly) excluded as possibility, since the algorithm must remain (partly) opaque. This 'fact' gives power to owners and developers of algorithms.

Apart from looking at the context of the algorithm, we need to look at the kind of discourse in which the algorithm is imbedded, to understand whether there are powerful agents that maintain the problem of opacity.

The second stage in the framework concerns identification of the problems that need to be tackled. This can be done through analysis of: a) "The network of practices it is located within", b) "The relationship of semiosis to other elements within the particular practice(s) concerned", c "the discourse (the semiosis itself)" (Fairclough 2001a, 125). I understand this stage as putting the texts into context (into the whole) in a heuristic way and then focus on its parts, by performing a semiotic analysis. The first two steps of stage two are part of contextualising the problem, as I have already done in earlier chapters: in chapter 1.1 "What is an algorithm? What is considered (part of) an algorithm?" I have elaborated on the actors that are present in the algorithmic network and how these are all intertwined. In chapter 1.2 "Current discussions on search algorithms" I have outlined how discussions on search algorithms touch upon other problems, such as bias and discrimination.

The last step in the second stage is the semiosis: the actual analysis of the texts itself. Whereas step one and two in this stage where looking at the texts within the whole (its context), this steps zooms in on the texts itself. This step consists of four parts. These are: structural analysis (the order of discourse), interactional analysis, interdiscursive analysis, and linguistic and semiotic analysis (Fairclough 2001a, 125). Structural analysis is concerned with how language is used in interaction to structure or influence, for example: "the way in which managerial discourse has colonized public service domains such as education" (Fairclough 2001a, 126). Interactional analysis consists of interdiscursive analysis and linguistic and semiotic analysis. In short: texts are interactive even if they are written—in which case the interactants are distant. Interdiscursive analysis covers analysing how different genres, discourse and styles are used together in articulations. Linguistic and semiotic analysis strongly rely on linguistic categories: "the core operationalizations depend on linguistic concepts such as actors, mode, time, tense, argumentation, and so on" (Meyer 2001, 25). This third step (the discourse itself) with its four parts are the core of my thesis and can be found in the next chapter (Chapter 3: "Analysis/ interpretation").

There is no clear distinction between the process of data collection and analysis: "CDA sees itself more in the tradition of Grounded Theory where data collection is not a phase that must be finished before analysis starts but might be a permanently ongoing procedure" (Meyer 2001, 18; Meyer 2001, 23–24). Also, CDA lacks a specific way of collecting data (Meyer 2001, 23). Therefore, the last step in the second stage (the step that consisted of the four aforementioned parts: structural analysis, interdiscursive analysis, linguistic and semiotic analysis) was not done in a specific order. Rather, as will be elaborated on in in the next chapter, during the research the analyses of the texts were constantly moving back and forth between these parts of the analysis.

The next stage of Fairclough's approach is stage 3: "consider whether the social order (network of practices) in a sense 'needs' the problem" Fairclough 2001a, 124–125). Essentially, this stage has already been covered in the introduction and first chapter, since this stage is about whether the problem is a constructed problem or not, for whom it is a problem, what the relevance is of addressing this problem. In the conclusion I will touch upon this stage once more in light of the findings of stage 2 and as part of stage 4, which is concerned with the identification of possible ways past the obstacles. The drawing up of stage 5, the critical reflection on the analyses of stages one to four is covered in chapter 4.

As mentioned before, CDA is not a strict approach. It is encouraged to create a process that suits a specific case. Therefore, I have not blindly followed Fairclough's list of steps and stages, but kept thinking how certain stages, steps and parts were productive for my research.

Since the stages, parts, and steps that were described in this section can be confusing, I have once again printed the overview of the framework below and included the chapters and of my thesis that address specific parts of the framework:

Stage 1	Focus upon a social problem that has a semiotic aspect.	Introduction Chapter 1.2: "Current discussions on search algorithms"
Stage 2	Identify obstacles to it being tackled, through analysis of a) the network of practices it is located within b) the relationship of semiosis to other elements within the particular practice(s) concerned	Chapter 1.1: "What is an algorithm? What is considered (part of) an algorithm?" Chapter 1.2: "Current discussions on search algorithms"
	c) the discourse (the semiosis itself) • structural analysis: the order of discourse • interactional analysis • interdiscursive analysis linguistic and semiotic analysis	Chapter 3: "Analysis/ interpretation"
Stage 3	Consider whether the social order (network of practices) in a sense 'needs' the problem.	Introduction Chapter 1.2: "Current discussions on search algorithms" Chapter 5.1: "Is the cure better than the disease?"
Stage 4	Identify possible ways past the obstacles	Chapter 5.2: "Possible ways past the obstacles and suggestions for further research"
Stage 5	Reflect critically on the analysis	Chapter 4: "Reflection and

limitations"

Table 1: Fairclough's stages and the chapters of this thesis in which the findings of these stages can be found.

In the next section, the specific procedure of data collection and data analysis that was used for this research is described.

2.2 Methods of data collection and analysis

2.2.1 Data, data collection and selection

For the analysis, the following texts (all websites) were used. The data can be divided into two parts. Texts that represent Google in general, and texts in which Google Search specifically is presented. Together, these make up the corpus that was used for this study:

- General representation of Google
    - Google "Ten things we know to be true" ("Ten Things We Know to Be True" 2015)
    - "Answers about Privacy and Security" ("Answers about Privacy and Security" 2015)
- Specific presentation of Google Search
    - Google Inside Search "Algorithms" ("Algorithms" 2015)
    - Google Inside Search "Crawling & Indexing" ("Crawling & Indexing" 2015)
    - Google Inside Search "Fighting Spam" ("Fighting Spam" 2015)
    - Google Inside Search "Policies" ("Policies" 2015)

The above corpus was combined manually. In first instance, it was intended to perform a quantitative comparative analysis in which Google's blog posts about their search algorithm were compared on word frequency and concordance with word frequencies and concordance in the National Bank of English (BoE) or Corpus of Contemporary English (COCA), both of which include lists of real word use of many genres.[13] However, it was found that Google has many different blogs on different topics while their tags are inconsequently employed. A blog that did discuss its algorithm was, for instance, found on their general company blog ("Our Secret Sauce" 2015), instead of their In Search blog (their blog about search). Because of the huge amount of blogs posts produced by Google on various topics, it was decided to take a different approach, since the collection and selection of texts would be too time consuming. For the eventual study, the texts on Google as a company, such as their philosophy, were used to study how Google linguistically presents itself. In this study, it is assumed that Google's general philosophy also applies to their products, including Google Search. Google's general self-presentation therefore also includes their other pages that are listed above in the corpus list.

[13] In a similar way as Tony Bastow performed his study on rhetoric in defence discourse (Bastow 2008).

2.2.2 Methods of data analysis

Quantitative data analysis

To avoid confirmation bias as much as possible,[14] and for the sake of discovery of unforeseen patterns, the academic corpus was, firstly, compared with the Google corpus on word frequency to discover differences in extensive use of certain words, or lack of use of words.[15] Since the process of collection, interpretation, and analysis is non-linear the results from this qualitative analysis were analysed, *before* going further to the next step in analysis, since the findings possibly influenced the direction of focus in the qualitative analysis. Only the 5000 most frequent words were showed in NVivo (*NVivo Qualitative Data Analysis Software* 2014)[16]. In this list, I looked for words (and their synonyms) that the academic literature had considered important, such as opacity, transparency, ethics, responsibility, bias and discrimination (see Table 2 in the next chapter).[17]

Some differences between the corpuses stood out (see Table 2, Figure 5, and Figure 6 in chapter 3: "Analysis/ interpretation"). Consequently, a qualitative approach was performed to understand how these words were used. While doing the qualitative analysis and the interpretation of the findings, the findings of the quantitative analysis were used to compare quantitative and qualitative findings: whether they reinforced or contradicted each other (see chapter 3: "Analysis/ interpretation") In chapter 4.2, "Using mixed methods", I will reflect on using both quantitative and qualitative analysis.

Qualitative analysis: close reading and semiotic analysis

The texts were subjected to multiple close readings and coding. The codes were distilled from Baker and Ellece's *Key Terms in Discourse Analysis* (2011) and include elements such as modality, the use of metaphors, and mitigating words (See Appendix A for the list with codes). The elements in this list were selected based on the criterion that these elements are often studied in Critical Discourse Analysis, since they may be indicators or unequal power relations, or rhetorical techniques.

---

[14] Combining different methods, such as quantitative analysis of corpuses and close qualitative analysis is a form of triangulation, which minimizes the risk of bias — Reisigl and Wodak propose that this multi-methodological approach should be done in combination with "a variety of different empirical data as well as background information" (Reisigl and Wodak 2001, 35).

[15] My approach could be regarded as a mix between a corpus-driven and corpus-based approach: in a corpus-driven approach minimal theoretical assumptions are used. It relies for its direction of analysis on "frequency and other statistical analysis", while the later uses "the corpus as a source of examples to check researcher intuition" (Baker and Ellece 2011, 29). This study seems somewhere in between: on the basis of researcher intuition, a corpus is put together and this is firstly subjected to an exploratory data analysis (corpus-driven), which directs in consequence the researcher's intuition (Baker and Ellece 2011, 29).

[16] While Nvivo is *qualitatitive* data analysis software, I have only used the *quantitative* features such as word frequency queries and lists.

[17] Baker and Ellece would possibly consider this approach to be subsumed under corpus-assisted discourse studies (CADS) (2011, 24–25).

After marking (coding) all these indicators of linguistic strategies, findings from the quantitative analysis were compared to findings from the qualitative analysis. For instance, if in the close reading analysis it was found that Google often uses the word "new" to describe its products, this finding was checked with findings from the quantitative analysis to, firstly, ensure that indeed "new" was used more often than other words (and this conclusion was thus not the result of confirmation bias), while, secondly, it could also be examined if and when other words were used a similar amount of times. This would help to interpret the meaning of an often-occurring word, such as "new".

During the close reading phase, words and phrases were merely coded, while during the interpretation phase, the findings were considered in context to gain deeper insights in the used language and its implications for maintaining, reinforcing, weakening or constructing power relations. I will discuss the findings and interpretation in the next chapter.

3 Analysis/ interpretation

To understand how the algorithm is being shaped by discourse, in this case Google's discourse specifically, a corpus of texts produced by Google was analysed: both quantitatively and qualitatively, as discussed in the previous chapter.

3.1 Agency, activeness, passiveness – relation between Google and user

It was found that Google's discursive construction of users' agency is ambiguous; on the one hand, users are often presented as passive social actors, having little to no agency, while simultaneously, user choice and thereby user agency is stressed. I will try to show these inconsistencies in the following examples.

Firstly, the mention of Google's users (most often referred to with "you" or "user") repeatedly coincides with words that render them as helpless, such as "protect" and "warn" while the user is linguistically passivized:

> This allows us to **protect** our users from… ("Fighting Spam" 2015)
> Learn more about how **we keep you safe** online ("Answers about Privacy and Security" 2015)
> we try to alert the site's owner **to help him or her**… ("Crawling & Indexing" 2015)
> we **warn you** when you try to… ("Answers about Privacy and Security" 2015)
> Google's … technology **protects** over 1 billion people, **warns** you when… ("Answers about Privacy and Security" 2015)
> to **help** people access and use even more… ("Ten Things We Know to Be True" 2015)
> When you use Google services, you are **protected** ("Answers about Privacy and Security" 2015)

According to Baker and Ellece, social actors "can be represented as 'doing' things (as actors/agents) or as having things done to them (as goals or beneficiaries of other social actor's actions)" (2011, 88). The former are active actors, which "make things happen and can therefore influence their environment", while the latter are passive and cannot have "meaningful influence on their environment" (Baker and Ellece 2011, 88). This passivization of users disempowers the user by constructing them as inactive, while empowering Google.

However, Google represents itself not only as powerful, but also as submissive by repeatedly using words and phrases that indicate obedience, thereby seemingly ascribing the user power/dominance:

> they [Google's products] will ultimately **serve** you ("Ten Things We Know to Be True" 2015)
> we're ultimately **serving** all our users ("Ten Things We Know to Be True" 2015)

Yet, this rendering of Google as servant may also be used as justification for something yet unknown. I will come back to this possibility at the end of this subchapter and try to understand what that something could be.

Secondly, while content wise the user is occasionally represented as having control, its linguistic representation renders the user as passive, since generally "we" or "Google" is the active actor, while the user is the beneficiary:

> "**We** . . . **give you tools** to control the types of data we collect and use" ("Answers about Privacy and Security" 2015)
> "**We** . . . **put you** in control" ("Answers about Privacy and Security" 2015)
> "We have . . . **tools that help you** control how Google works for you" ("Answers about Privacy and Security" 2015)

Again, the user is a passive actor, while Google is the active actor. Thereby, the content of what Google claims (putting the user in control) is contradicted by its linguistic form. Thus, while on the surface level the user is empowered, covertly, the user is disempowered.

Thirdly, even when the user is the active agent in a sentence, often the user's agency is related to, or enabled by, the use of Google's products:

> **Using our translation tools**, people can discover content… ("Ten Things We Know to Be True" 2015)
> the actions you take **using our services**… ("Answers about Privacy and Security" 2015)
> When you **use our services** ("Answers about Privacy and Security" 2015)
> what you create **using our services** ("Answers about Privacy and Security" 2015)
> **With our Ads Settings tools**, you can control ads ("Answers about Privacy and Security" 2015)

It seems that while linguistically the user is an active agent, its agency is mitigated by the content of Google's articulations, which stress users' dependence on Google's products. Additionally, when the user is the agent in an active sentence, often the user's agency is rendered by negation:

> "**you don't have to** consider" ("Ten Things We Know to Be True" 2015)
> "**you don't need to** be" ("Ten Things We Know to Be True" 2015)
> Even if **you don't know** exactly what you're looking for… ("Ten Things We Know to Be True" 2015)

While the user is an active agent in these sentences, his or her influence or actions are described in terms of what the user does *not* (need to) do, stressing the *not* taking of actions. In other words: while generally activated social actors "make things happen" (Baker and Ellece 2011, 88), presenting users' actions in negating form undermines the activation of users.

While the above examples show that the user is discursively constructed as disempowered actor, either linguistically or substantively, the user is sometimes put in power, both substantively and linguistically (as an active agent). Firstly, the user is empowered by the use of the modal verb *can:*

> you **can** find tools ("Answers about Privacy and Security" 2015)
> you **can** manage ("Answers about Privacy and Security" 2015)
> you **can** control ("Answers about Privacy and Security" 2015)
> people **can** discover ("Ten Things We Know to Be True" 2015)

"Can" is a modal verb that expresses possibility (epistemic modality) (Baker and Ellece 2011, 71). In these examples, *can* stresses user choice, since *can* allows the user to decide whether to manage, or not manage, to control, or not control. According to Paul Baker, "relatively powerful groups seem to be paired with modal verbs which give them more freedom and choice," such as *might* and *can*, "while more controlling modal verbs," such as *should*, "are used with less powerful groups" (Baker 2006, 160). It thus seems that the user is discursively ascribed power in two ways: by being an active actor, and by the use of model verbs denoting possibility. Yet, I would argue that possibility also expresses uncertainty, undermining user agency: "you manage" and "you control" sound stronger than "you can manage" or "you can control".

In conclusion, it seems that the discursive construction of user agency is ambiguous. The user is presented as a beneficiary, a passive actor, and/or as in need of Google's products to obtain power, while at the same time the user is discursively constructed as having choice, as actors that should be served, which ascribes them power.

Also, it seems that the relation between Google and its users is complex. Both are discursively ascribed with power. If we step away from the text and look at the context, it seems that the relationships between Google and its users can be described as relationships of interdependence. Users trust Google and are dependent on Google; yet, Google is also dependent on her users. This complex relation of power is reflected in the ambiguous discursive construction of user agency *and* the discursive construction of Google itself. While Google represents itself, as I have argued earlier, as powerful, they render themself also as submissive: serving the user etc. This may also be a rhetorical technique to reassure the user of Google's good intentions; again, this only stresses the complex mutual dependency of Google and her users.

3.2 Absence replaced by the technical imaginary

From the former section, it may be clear that Google expresses to be concerned with its users, that Google takes "great care" to "serve" her users ("Ten Things We Know to Be True" 2015); yet, there is near to total absence of topics of transparency, discrimination, or bias—topics that are considered important by scholars and media critics and concern the user. In the preliminary, exploratory word frequency analysis, it was found that topics that were key in recent discussions in academic literature were not, or barely, mentioned by Google (Table 2, Figure 5, Figure 6):

	Frequency in academic corpus		Frequency in Google corpus	
Keyword	Weighed percentage	Count	Weighed percentage	Count
bias	0,14%	205	-	-
transparency	0,08%	118	0,08%	4
discriminate	0,08%	113	-	-
discriminatory	0,01%	9	-	-
disclosure	0,02%	36	-	-
disclose	0,02%	27	0,03%	1
Regulation	0,06%	88	-	-
law	0,16%	237	0,03%	1
Rules	0,15%	226	-	-
Ethic	0,49%	736	-	-
Value	0,16%	246	0,08%	3
norm	0,02%	31	-	-
open	0,05%	78	0,10%	4
Opacity	0,01%	9	-	-
opaque	0,01%	15	-	-
Neutral	0,01%	19	-	-
Spam	0,01%	10	1,23%	49
Abuse	0,01%	8	0,05%	2
Choice	0,06%	89	0,10%	4
Choose	0,03%	43	0,08%	3
Users	0,26%	395	0,45%	18
Need	0,23%	343	0,20%	8
want	0,07%	101	0,28%	11
Require	0,14%	216	0,08%	3
request	0,02	31	0,48%	19
Guidelines			0,23%	9
Hacked			0,28%	8
Relevant			0,35%	14
Responsibility			-	-

Table 2: Words in this table were manually chosen from the topics that were addressed in the academic literature. For instance, from reading the academic corpus closely, it was found that "opacity" and "users" were often addressed. Therefore, they were included in the list. On the other hand, it was found that words such as "spam" were often used by Google, which is why these words were also included in the topic list. In the word frequency analysis, the frequencies of these words were looked up using NVivo. There is no specific order in this table.

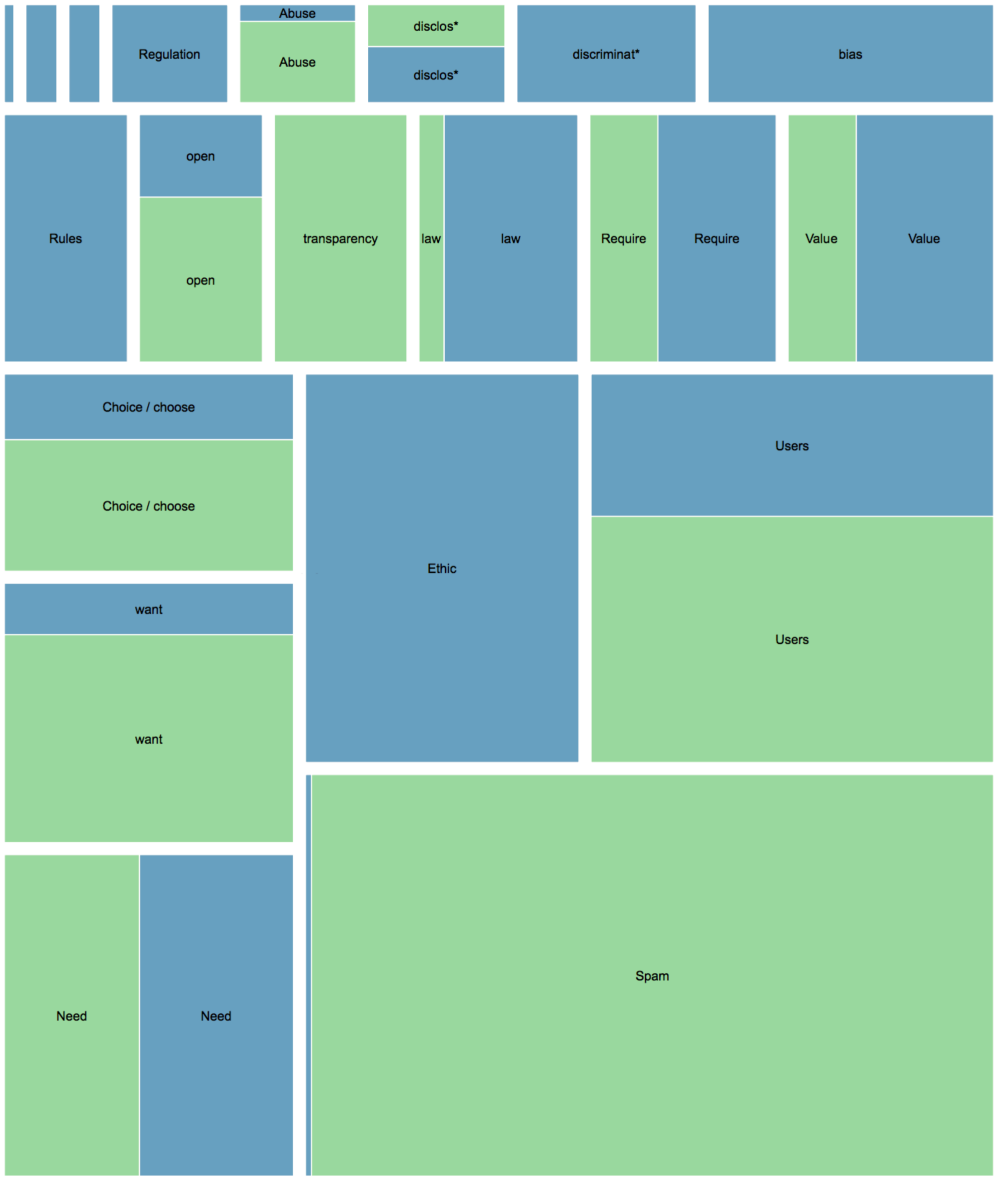

Figure 5: Treemap visualization of the data in table 2, made with Raw (Caviglia et al., n.d.). The sizes of the boxes present the weight percentage of the designated word. The green visualizes the weighed percentage in the Google corpus, while the blue parts indicate the weighed percentage of the designated word in the academic corpus.

As can be seen in Table 2, there is no explicit mention of bias, discrimination, ethics or responsibility in the Google corpus. Google does not mention opacity, however, transparency *is* mentioned. It is used in sentences that refer to Google's transparency report, which I decided to not include in the Google corpus for this study, since it focuses on what Google does or not do with its data, such as giving user information to governments ("Access to Information" 2015), while the focus of this study is rather the discourse on transparency of the workings of the algorithm. While Google does not go into issues of transparency in relation to the workings of its algorithm, Google is transparent, however superfluously, about properties about and processes that concern its search algorithm, such as that the Google search algorithm uses more than 200 signals ("Algorithms" 2015) and

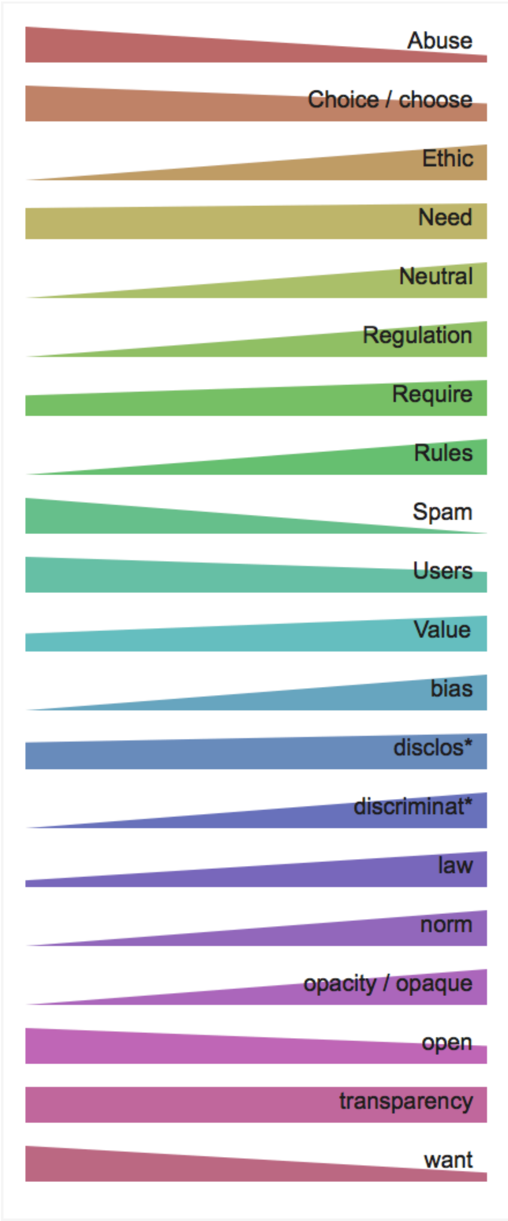

Figure 6: Small Multiples visualisation of the word frequency in the Google corpus (left) versus the word frequency in the academic corpus. For each word, the frequency in the Google corpus is compared to the frequency in the academic corpus. For each word an independent scale is used. For this visualisation, RAW was used (Caviglia et al., n.d.)

that content from search results can be removed under certain conditions ("Policies" 2015). Also, Google does indicate that sponsored links "are *clearly* marked as such" ("Ten Things We Know to Be True" 2015), but offers no insight in the workings of the algorithms. Transparency as discussed by scholars, namely as giving full disclosure of the algorithmic code, its data and the criteria that are used for the data, is absent from Google's discourse.

I would argue that this gap, so it seems, is rather replaced by utopian promises, thereby distracting the reader from recent problems and discussions concerning ethics, bias, discrimination, and transparency. The Google corpus is dense with references of products and services that are yet to be build. The word *new* for instance, is one of the most occurring words in the Google corpus (the tenth most frequent, to be exact).[18] Moreover, many products or processes are described in terms of *better* and *faster*, for instance; but what does *better* means, and for whom is it better? What is *new*, is it really new, or is *new* only used to capitalize on user's hopes? Google does not, or barely elaborates on these concepts, but merely uses them. By close reading the Google corpus again, it was attempted to find answers to how Google exactly shapes this utopian replacement and how this replacement contributes to the discursive construction to Google's Search Algorithm.

It was found from the close reading analysis that using Google's products, including search, should be a "seamless experience" ("Ten Things We Know to Be True" 2015), it should be "fast" ("Ten Things We Know to Be True" 2015), "free" ("Ten Things We Know to

[18] The eight most frequent words are all words that are either a synonym for *search algorithm*, or an aspect of it ("google", "results", "content").

Be True" 2015), and the search results should be "relevant" ("Ten Things We Know to Be True" 2015; "Answers about Privacy and Security" 2015; "Fighting Spam" 2015; "Policies" 2015). Apparently, these properties are common ideals: they are not explained (what is relevance, for whom is it relevant, why is it relevant (see also Gillespie 2014, 127), or they are repeatedly used as an unexplained argument or mentioned as an important asset, such as the word "free", as if free products or services are an ultimate goal:

> Data also helps us show relevant ads, so we can make our services **free** for everyone. ("Answers about Privacy and Security" 2015)

> We also use data to show you ads that are relevant and useful, and to keep our services **free** for everyone. ("Answers about Privacy and Security" 2015)

> Ads are what enable us to make our services like Search, Gmail, and Maps **free** for everyone. ("Answers about Privacy and Security" 2015)

> In addition, we're hoping to fuel greater innovation for mobile users everywhere with Android, a **free**, open source mobile platform. ("Ten Things We Know to Be True" 2015)

While "free", for instance, does not occur that often as words such as "new", it does add up to the creation of some mythical things that we are striving towards. Seamless communication reinforces this idea. It also seems a term that refers to some common ideal. In fact, the idea of seamless communication is not new at all (see for instance: de Vries 2013). Due to time and space restrictions I cannot go further into this interesting topic. The point that I want to make is that the Google corpus has on the one hand gaps with regard to topics that have been addressed as very important, while on the other hand the corpus is full of references to some ideal future. The latter seems to function, whether intentionally or not, as a veil for the former: readers may not even be aware of this gap, because their attention is led elsewhere.

3.3 The representation of the algorithm as a manual and automatic instance

As argued in chapter 0, "Defining the search algorithm in context of current debates", the algorithm can be considered to be the code itself, but it can also be considered the whole network of its actors, including its developers and users. In the close reading analysis, it was found that Google is very explicit in stating which processes include automatic processes and which are processes that are performed manually. This finding raises questions such as: why does Google distinguish in such an explicit way between automatic and manual? What are implications of discursively constructing these as two different processes?

It was found that Google implicitly contradicts her own explicit claims about automatic processes and manual processes. On the "Policies" page it is stated in a heading: "Algorithms Over Manual Actions" with the elucidation: "We prefer machine solutions to manually organizing information" ("Policies" 2015). While one would expect to be more deeply informed about the reasons behind this preference, or what those machine solutions might be, the focus in the Google corpus is rather about Google's *manual* actions. This claim seems substantiated by findings from the quantitative analysis: "manual" and its stemmed words occur 24 times in the GC, while "automatic" and its stemmed words occur four times. Word combinations including the word "machine" occur twice (which is twice in the combination "machine solutions").[19] Of these four occurrences of "automatic" only one actually concerns Google's search algorithm. The other three refer to the automatic content creation of spammy sites. In the final chapter I will discuss to which extent conclusions can be drawn on basis of frequency analysis and, more generally, how quantitative methods relate to qualitative methods for the interpretation of data in a mixed methods study such as this one.

Moreover, it was found that when automatic algorithmic actions are discussed (without using explicit words such as "automatically") positive algorithmic actions are often expressed by modals that express possibility and render the algorithm powerful, as described in the former section, while negative outcomes of algorithmic actions are mitigated:

> in some cases, our algorithms **falsely identify** sites … ("Policies" 2015)

> When one of these algorithms **misidentifies** websites (for example essex.edu) we sometimes make manual exceptions to prevent these sites from being **classified** as pornography" ("Policies" 2015).

What is implicitly stated here is that the algorithm can be wrong, or make mistakes. However, mitigating words are used that render these mistakes less severe: "misidentify" or "falsely identify" are weaker. Additionally, the deletion of an agent in the second example ("to prevent these sites from being classified") can make the agent appear to be "discursively absolved from responsibility", as is often the case if agents are deleted in passivized sentences (Baker and Ellece 2011, 87).

To sum up: algorithmic actions are thus not explicitly discussed as being automatic. Google rather focuses on elaborating on manual actions, which may be because a focus on manual action may be perceived as more humane and therefore be more agreeable. Moreover, negative outcomes of algorithmic actions are mitigated, while actions of spammy sites are disproportionally enlarged. The latter will be focused on next.

3.4 We, Google, versus them, hackers.

One of this thesis' subquestions is how the algorithm is discursively is constructed by Google with regard to spammers. In the academic literature, most scholars propose more transparency of the algorithmic code, of the data that is used by the algorithms and the criteria that are used. Yet, they acknowledge that full disclosure will be harmful, allowing the

---

[19] This word was also searched for, since word combinations as "machine processes" or "machine solutions" can be a substitution for "automatically".

misuse by hackers etc. Does Google play a part of this construction of them (hackers) versus us (Google, users), and if so, how? To gain more insight, the qualitative analysis also focused on how Google was described in relation to how spammers and hackers were discursively rendered.

It was found that aside of stressing the 'human side' of the algorithm (as argued in the previous section) Google presents itself as caring, using words that appeal to emotions and intensifying words such as "hard" and "deeply":

> **We try hard** to make information available … ("Policies" 2015)
> **We care deeply** about the information you find on Google. ("Policies" 2015)
> **We'd love** to get your feedback … ("Policies" 2015)
> **We hate** spam as much as you do … ("Policies" 2015)
> **We want to be careful** … ("Policies" 2015)
> **Fighting** spam ("Policies" 2015)

Moreover, as discussed in a previous section, Google also renders herself as a protector, by using words and phrases such as "we keep you safe" ("Answers about Privacy and Security" 2015). In contrast, spammy sites are discursively constructed as some sort of enemy. Google labels spammy sites as "dangerous", that can "hurt" the user ("Policies" 2015), that cause relevant websites to "get buried", and are "aggressive" ("Fighting Spam" 2015); therefore Google needs to "fight" spam ("Policies" 2015). Fighting is generally associated with war and the enemy. Also the former words have connotations of violence. As a consequence, the distance between Google as good and hackers/spammers as bad increases, resulting in reinforcement of Google's positive, protecting intentions, while the negative properties of this unknown 'other' are magnified. This 'othering' is reinforced by stressing the difference between "them" and "us", for instance, by repeatedly referring to the search results as "our search results", while in the rest of the corpus this is not done, except for one case.[20]

> remove **them** from **our** results. ("Policies" 2015)
> Sometimes we remove content or features from **our** search results for legal reasons ("Policies" 2015)
> We also disclose certain details about legal removals from **our** search results through **our** Transparency Report. ("Policies" 2015)

Possibly, "our" is used to strengthen Google's argument. Removal is a sensitive topic—removal might lead to or be a result of censorship—and by using "our search results"

[20] This other occurrence of "our results" refers to Google's policy on advertisements: "Advertising on Google is always clearly identified as a 'Sponsored Link,' so it does not compromise
the integrity of our search results. We never manipulate rankings to put our partners higher in our search results and no one can buy better PageRank" ("Ten Things We Know to Be True" 2015). Also in this case, it seems that "our" is added to reinforce Google's statement that concerns a controversial issue, namely sponsored links. By using "our" the user may more easily accept Google's policies, since it is *their* product.

instead of "the search results" or "search results" it is stressed that the removal of results concerns Google's *own* product, which might renders removal more acceptable or natural and thereby more agreeable.

It seems that Google attempts to position the hacker or spammer as the dangerous other. Google may be perfectly right; there is much spam and malware on the Internet. In this case however, it seems that focusing on misuse of the 'enemy' is used as a rhetoric strategy to win the reader. Google's good intentions are reinforced; by rendering the other as dangerous, the self is considered less dangerous.

3.5 Technologic evolution as inescapable and natural

Epstein notes in his article how Google uses the word 'organic'. He claims the following:

> Google's search algorithm is pushing one candidate on the top of rankings because of what the company coyly dismisses as 'organic' search activity by users; it's harmless, you see, because it's all natural. Under this scenario, a computer program is picking our elected officials. (Epstein 2015).

Epstein does not refer to the source in which Google speaks of 'organic' search activity. However, since other powerful companies do speak about 'organic' traffic, such as Facebook's dashboards for page owners, it was decided to investigate if, and how Google used words such as 'organic' to naturalize certain algorithmic actions, since in other fields it has been proven that labels that refer to nature are considered harmless.[21]

It was found that Google does indeed use words that allude to natural events. Spam is represented as having, for instance, "unnatural" links, while what is considered "unnatural" is not explained ("Fighting Spam" 2015). To explain why it is normal that some sites have less visitors than usual, Google states that "sites are experiencing the natural eb and flow of online traffic" ("Fighting Spam" 2015), thereby rendering digital processes as something natural, but also something that is beyond the control of humans (as is eb and flow). Also, Google uses names of animal species for her algorithms and algorithmic updates, such as "the Panda algorithm" and the "Penguin" code update ("Fighting Spam" 2015), which renders something abstract and unknown (an algorithm) into something we do know (harmless, even protected, animals). Moreover, Google speaks of "The evolution of search" and "how search has evolved" ("Algorithms" 2015), thereby presenting adaptations of the search algorithm as natural and inevitable—as a natural consequence of its past and current forms. By representing search as an evolving entity, the first steps are already taken and we have accepted it, which makes it more difficult to interfere with it.

Although Google uses words that refer to nature or natural processes, the Google corpus is not very dense with it. I would say that the use of these words might reinforce the discourse

[21] The food industry, for instance, uses labels such as "natural" and "pure" because users expect that such products are free from synthetic additives or manipulations, while this is often not the case.

that is constructed by Google, but does not play such a major role in it as other aspects do, such as absence and passivization of the user.

Nevertheless, it seems that in the discourse on algorithms, there is indeed a tendency to speak of the algorithm in terms of “organic”. Consider Granka’s suggestion about regulating algorithms: “If any aspect of search engine algorithms were to be regulated, the most important part is identifying when the search engine deviates from their *organic* algorithm to instead promote profit-making content” (Granka 2010, 369, my emphasis). Yet, she neglects explain what *organic* means: when is an algorithm organic? Without clear definition and criteria that constitute a norm, it is impossible to determine deviation from the norm. Thereby, Granka’s mentioning of “organic” seems to refer to some common knowledge about what organic is.

4 Reflection and limitations

This chapter covers stage 5 of Bhaskar's framework—which was discussed in chapter 2.1.2—and concerns a critical reflection on the analysis. I will firstly reflect on my research and discuss possible limitations to my research. Secondly, I will discuss how combining quantitative and qualitative analysis worked together for this specific study and in general.

4.1 Reflections and limitations

In this study it was found that Google uses specific words that render themselves as agreeable and that shift the user's focus to promising future possibilities instead of uncomfortable topics such as discrimination. However, there are some limitations to these findings.

Firstly, the Google corpus that was put together for this study is relatively small. While more is not necessary better, it is also not the case that less is always more. It was difficult to find texts produced by Google that specifically discuss its algorithm or its search engine, which is the reason why the corpus is relatively small. Yet, performing this study's analysis on a larger corpus of texts may lead to more substantiated findings. The use of words such as "unnatural" may now be coincidental and only have been used in a specific text.

Secondly, a comparative analysis between the Google corpus and a reference corpus may give more insight in contemporary word use and similarities or disparities in Google's corpus. Describing technology in words that allude to natural phenomena may, for instance, be not limited to Google's discourse, but can be part of a discourse in which the Google discourse is embedded. While the findings of this study would in that case still hold (using 'organic' may render a technology more agreeable), it would be clearer whether such a discourse is constructed by Google, or whether it is part of a larger discourse—resulting in different questions of power and responsibility.

Thirdly, the qualitative analysis may have been subject to confirmation bias. While I have attempted to prevent this as much as possible by performing both a qualitative *and* quantitative analysis, confirmation bias may nevertheless have crept into my analysis. I may too easily have ignored linguistic patterns and words that disagree with my hypotheses. Yet, by clearly describing the data that I have used, and the steps that I have taken in my analysis, other research can easily check whether they do or do not find the same.

Using this specific methodology has proven productive. Performing a 'normal' reading of the text for the first time resulted in a different interpretation than using close reading to find linguistic patterns and words that may indicate rhetorics, assumptions, linguistic strategies etc. A surface reading would allow me to only read the explicit text, while close reading enabled me to look beneath the surface. In fact, it is exactly at the point where the explicit reading is in tension with implicit readings that discursive strategies become visible.

Lastly, in chapter 1 it was discussed that the filter bubble effect amplifies confirmation bias. Since my research depended on Google Scholar to find academic articles and other information, my research process might have been subject to the filter bubble effect; that is, I might have been shown information that the algorithm thought I was looking for, while other articles or websites have been kept hidden. Moreover, since I do not know if and how the Google Scholar search algorithm works, I cannot determine whether this is the case, and if so, the implications for my research. While I did use various search engines to overcome the possibility of filter bubble effects, my research might have been affected. More generally, this notion of filter bubble effects raises questions such as what implications there are for academic research.

4.2 Using mixed methods

In this study both quantitative analysis and qualitative analysis were used. Quantitative analysis on its own, such as word frequency analysis does not necessarily show what is important. Topics can be addressed without explicitly naming them. As a matter of fact, especially words that relate to ideological struggles have many (near) synonyms (Baker and Ellece 2011, 85).[22] On the other hand, words that occur often are not necessarily important topics.

In this specific study quantitative analysis and qualitative analysis were combined and the outcomes were constantly compared with each other. Did these methods strengthen or weaken each other?

The methods complemented each other. Quantitative analysis was use to check preliminary findings from the qualitative analysis. For instance, it was noted that 'new' was mentioned a couple of times and that the use of 'new' occasionally seemed out of context. Using quantitative analysis (a word frequency list of the occurrences of every word in the Google corpus) it was found that 'new' was used many times, and my notion of the word 'new' was not due to confirmation bias. In this sense, quantitative analysis complemented qualitative analysis. It allowed to quickly check how often and in which cases a certain word was used, without the need to manual reread all the texts in the corpus.

Yet, a high word frequency does not say anything in itself. Qualitative analysis was needed to understand how a word such as 'new' was used: in what grammatical context and semantic context, in combination with which words, etc. It was also analysed which contextual elements were *absent*: thus, certain words were interpreted in relation to words that were *not* explicitly mentioned (such as bias and discrimination) to understand a word's function in the text.

Generally, this use of mixed methods is a form of triangulation. Its "term originates from geometry and land surveying, where an accurate view is obtained from looking at things

---

[22] In such a case, each word (such as 'terrorist', 'rebel', freedom fighter', 'assassin') may occur just once, while the subject that they relate to is mentioned more than once.

from two or more positions” (Baker and Ellece 2011, 154). Triangulation in academics entails using multiple methods for data collection or analysis. It allows the researcher to look at the research subject from multiple points of view (Baker and Ellece 2011, 154), it “facilitates validity checks of hypotheses, it anchors findings in more robust interpretations and explanations, and it allows the researcher to respond flexibly to unforeseen problems and aspects of the research” (Layder paraphrased in Baker and Ellece 2011, 154).

The power, for my research, was in combining these two methods, which indeed allowed me to do the above. From qualitative analysis hypothesis were formed, by quantitative analysis these could be checked and findings from the quantitative analysis were a source of possible other patterns or signs that could be studied in the qualitative analysis.

The qualitative and quantitative methods thus complemented each other. In the example of ‘new’, findings from the quantitative analysis strengthened and confirmed findings from the qualitative analysis. However, there were also cases in which findings form the quantitative analysis contrasted hypothesises that were derived from the close reading analysis. That is, in the qualitative analysis some words stood out, but they did not occur that often in the texts. Yet, also in these cases these methods complement each other: it allowed me to know, for instance, that a certain word stands out in a qualitative analysis, but that it is also used sparingly. This is also valuable information. My ultimate findings are thus substantiated by both quantitative and qualitative analysis.

While my corpuses were not very large, other studies that study large corpuses may especially benefit from using mixed methods to allow the researcher to analyse large corpuses.

5 Conclusion: from research findings to future research

5.1 Is the cure better than the disease?

According to Langdon Winner (1980), there is a certain discourse that justifies certain actions, often based on assumptions about how certain processes should be run:

> Whatever claims one may wish to make on behalf of liberty, justice, or equality can be immediately neutralized when confronted with arguments to the effect: 'Fine, but that's no way to run a railroad' (or steel mill, or airline, or communications system, and so on). In many instances, to say that some technologies are inherently political is to say that certain widely accepted reasons of practical necessity—especially the need to maintain crucial technological systems as smoothly working entities—have tended to eclipse other sorts of moral and political reasoning. (Winner 1980, 133).

Winner's example is reminiscent of discussions on transparency of algorithms when he refers to claims of *necessary* actions or policies. At this moment, we are at the point that powerful companies seem to claim: "Fine, we understand that you want to have transparency on the algorithm, but that is no way to run an ICT company". It is understandable that any company wants to control how they do their work, yet, we should be aware of too easily accepting these claims. Is there really a "*need* to maintain" algorithmic opacity and "what other sorts of moral and political reasoning"—to use Winner's words—are eclipsed if we (blindly) accept these claims?

Keeping algorithms opaque may lead to many unintended and unwanted issues, such as bias and discrimination. While algorithmic transparency, or rather the lack thereof, is more and more often addressed both in academics and popular media, we must not forget to scrutinize common starting points of these discussions and underlying beliefs. My study was initiated by questioning such a seeming common belief—the general idea that algorithms should (partly) be kept private to prevent spammers and hackers to abuse the disclosed information.

While my study has not found whether this claim is indeed true—which was not the purpose of this thesis—it did find that in the Google corpus various discursive strategies are used to render Google and her products in a certain way. These discursive renderings of the algorithm, its owner (Google), and its users may influence power relations between the latter two. To summarize, it was found that Google represents itself and its products including search as caring, while spammers are rendered as enemies by the use of words that are associated with war or conflict. Linguistic choices such as these may have justifying effects on Google's actions and make their policies more agreeable: we want to 'freely' use Google's products and we do not want hackers to spam our search results, so of course we agree with Google's policies that will ensure that Google's products are kept free—both free in the financial sense and free from spam and other 'bad things'. Yet, the idea that users want products to be free may be an idea that is constructed and not something that users really want, especially if it will be disclosed what the costs are of using free products—no attention is given to the costs however, only the benefits. Moreover, if we look at the algorithm as it relates to its users, it is explicitly constructed as an empowering device, while on the same time stressing the user's dependency. While on the surface Google seems to be cooperative, caring, and protective, on beneath this surface Google attempts to construct and control the needs of the user and thereby the user itself.

Topics can be addressed without explicitly naming them. Especially ideological topics are not explicitly mentioned, since they are common sense and do not need to be explained or mentioned; yet, not mentioning them does not mean that they are not important. For instance, if it is stated that the husband was bad because he was intimate with a women other than his wife, this judgement is probably based on the underlying ideology that one should not have sex outside marriage. Yet, it is not explained why out of marriage sex is bad. The ideology itself is not mentioned and may thereby be accepted without question. Similarly, if Google claims that disclosure of the algorithm leads to abuse by spammers, this claim also seems to lean on an underlying belief: such as that spammers are bad, and that codes and scripts should be kept away from them. In the Google corpus it was not mentioned *why* spammers are bad, similarly to as why having sex outside marriage is bad. In fact, Google seems to reinforce the idea that there are Internet villains out there from whom the user needs Google's protection.

What this thesis adds to current debates on algorithms is an awareness of covert discursive constructions that can construct or reinforce common beliefs, thereby manipulating users and also scholars, who are not aware of covert discursive mechanisms. It seems that in discursive constructions of the algorithm certain ideas are presented in such a way that they might reinforce or even co-constitute common assumptions, such as that spammers are enemies, that the user needs Google's protection and that keeping the algorithm is such a necessary way of protection. On the surface layer these discursive constructions are not visible. They seem to operate on a deeper level, which allows them to be passed unnoticed and unquestioned. We should be aware of such practices. Is disclosure really that bad? And what does Google ask in return for their protection? Is the cure better than the disease? In fact, is there really a disease?

5.2 Possible ways past the obstacles and suggestions for further research

Firstly, I would suggest that there is much yet to be researched when it comes to algorithms. From the literature review it was found that benchmarks are needed to find standards by which issues such as bias can be measured, but also standards for audits are needed. Defining benchmarks or norms asks for a multidisciplinary approach in which academics, governments and other parties need to work together, since algorithmic bias and other issues concerning algorithms address matters that concern academics, regulation, law and ethics.

Secondly, powerful companies such as Google need to use language that contains the same explicit as implicit meaning. This means that if there are general concerns regarding algorithmic transparency (or the lack thereof) it is not enough to make a transparency report in which it is discussed when search results are removed, while not mentioning transparency of the algorithm itself. Media critics, journalists and scholars have a responsibility of addressing such cases. The role of governments and regulating actors is yet to be determined.

Thirdly, researchers should be aware of discursive mechanisms, such as Google's discursive practices that have been addressed in this thesis. Similarly to how media critics may pay attention to framing by politicians, scholars should be aware of discursive constructions of algorithms by software companies. Underlying assumptions should be brought to the surface so that any ungrounded assumptions can be addressed and, if necessary, changed.

It may be clear by now that there are many issues raised regarding the intensive use of algorithms in society. Further research may extend on the research done in this thesis by incorporating other powerful agencies, such as Facebook to (better) understand how ideas are naturalized and what is veiled behind ideological promises. What claims are made and on what assumptions are these based? Are these assumptions grounded? What are users expected to give in return?

6 Bibliography

"Access to Information." 2015. *Google Transparency Report*. Accessed November 1. http://www.google.com/transparencyreport/.

"Algorithm." 2010. *Concise Medical Dictionary*. Oxford: Oxford University Press. http://www.oxfordreference.com/view/10.1093/acref/9780199557141.001.0001/acref-9780199557141-e-282.

"Algorithms." 2015. *Google Inside Search*. Accessed October 22. https://www.google.com/insidesearch/howsearchworks/algorithms.html.

"Answers about Privacy and Security." 2015. *Google - Answers about Privacy and Security*. Accessed October 14. https://privacy.google.com/.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Publishing.

Baker, Paul, and Sibonile Ellece. 2011. *Key Terms in Discourse Analysis*. London: Continuum.

Baker, Paul, and Amanda Potts. 2013. "'Why Do White People Have Thin Lips?' Google and the Perpetuation of Stereotypes via Auto-Complete Search Forms." *Critical Discourse Studies* 10 (2): 187–204. doi:10.1080/17405904.2012.744320.

Bastow, Tony. 2008. "Defense Discourse II: A Corpus Perspective on Routine and Rhetoric in Defence Discourse." In *Language and Power: An Introduction to Institutional Discourse*, edited by Andrea Mayr, 138–62. London: Continuum.

Breeze, Ruth. 2011. "Critical Discourse Analysis and Its Critics." *Pragmatics* 21 (4): 493–525.

Caviglia, Giorgio, Giorgio Uboldi, Matteo Azzi, and Michele Mauri. n.d. *RAW*. Density Design. http://raw.densitydesign.org.

"Crawling & Indexing." 2015. *Google Inside Search*. Accessed October 22. https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. "Automated Experiments on Ad Privacy Settings." *Proceedings on Privacy Enhancing Technologies* 2015 (1): 92–112.

de Vries, Imar. 2013. *Tantalisingly Close: An Archaeology of Communication Desires in Discourses of Mobile Wireless Media*. Amsterdam: Amsterdam University Press. http://www.press.uchicago.edu/ucp/books/book/distributed/T/bo13245934.html.

Diakopoulos, Nicholas. 2014. "Algorithmic Accountability Reporting: On the Investigation of Black Boxes." *Tow Center for Digital Journalism, Columbia University*. http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. "Fairness Through Awareness." *arXiv:1104.3913 [cs]*, April. http://arxiv.org/abs/1104.3913.

Epstein, Robert. 2015. "How Google Could Rig the 2016 Election." *POLITICO Magazine*. August 19. http://www.politico.com/magazine/story/2015/08/how-google-could-rig-the-2016-election-121548.

Epstein, Robert, and Ronald E. Robertson. 2015. “The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections.” *Proceedings of the National Academy of Sciences* 112 (33): E4512–21. doi:10.1073/pnas.1419828112.

Fairclough, Norman. 1989. *Language and Power*. Language in Social Life Series. Essex: Longman.

———. 1992. *Discourse and Social Change*. Polity Press.

———. 2001a. “Critical Discourse Analysis as a Method in Social Scientific Research.” In *Methods of Critical Discourse Analysis*, edited by Ruth Wodak and Michael Meyer, 121–38. Introducing Qualitative Methods. London: SAGE.

———. 2001b. *Language and Power*. Language in Social Life Series. Longman. https://books.google.be/books?id=ZRHCNMN3qqUC.

“Fighting Spam.” 2015. *Google Inside Search*. Accessed October 22. https://www.google.com/insidesearch/howsearchworks/fighting-spam.html.

Foucault, Michel. 1972. *The Archaeology of Knowledge*. Translated by Sheridan Smith. 2nd ed. Routledge Classics. Routledge.

Frosh, Paul. 2001. “Inside the Image Factory: Stock Photography and Cultural Production.” *Media, Culture & Society* 23 (5): 625–46. doi:10.1177/016344301023005005.

Gillespie, Tarleton. 2014. “The Relevance of Algorithms.” In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot, 167–93.

Granka, Laura A. 2010. “The Politics of Search: A Decade Retrospective.” *The Information Society* 26 (5): 364–74. doi:10.1080/01972243.2010.511560.

Hamilton, Kevin, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. “A Path to Understanding the Effects of Algorithm Awareness.” In *CHI’14 Extended Abstracts on Human Factors in Computing Systems*, 631–42. ACM. http://social.cs.uiuc.edu/papers/pdfs/paper188.pdf.

Hannak, Aniko, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. “Measuring Personalization of Web Search.” In *Proceedings of the 22nd International Conference on World Wide Web*, 527–38. International World Wide Web Conferences Steering Committee. http://personalization.ccs.neu.edu/papers/web_search.pdf.

Hinman, Lawrence M. 2005. “Esse Est Indicato in Google: Ethical and Political Issues in Search Engines.” *International Review of Information Ethics* 3 (6): 19–25.

Ince, Darrel, ed. 2009. “Algorithm.” *A Dictionary of the Internet*. Oxford: Oxford University Press. http://www.oxfordreference.com/view/10.1093/acref/9780199571444.001.0001/acref-9780199571444-e-88.

Introna, Lucas, and Helen Nissenbaum. 1999. “Sustaining the Public Good Vision of the Internet: The Politics of Search Engines.” *Working Paper* 9. https://www.princeton.edu/culturalpolicy/workpap/WP09%20-%20Introna+Nissenbaum.pdf.

Kay, Matthew, Cynthia Matuszek, and Sean A Munson. 2015. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations.” In *Proceedings of the*

33rd Annual ACM Conference on Human Factors in Computing Systems, 3819–28. ACM.
Kirchner, Lauren. 2015. "When Discrimination Is Baked Into Algorithms." *The Atlantic*, September 6. http://www.theatlantic.com/business/archive/2015/09/discrimination-algorithms-disparate-impact/403969/.
Kun, Jeremy. 2015. "Big Data Algorithms Can Discriminate, and It's Not Clear What to Do about It." *The Conversation*. August 13. http://theconversation.com/big-data-algorithms-can-discriminate-and-its-not-clear-what-to-do-about-it-45849.
Lister, Martin, Jon Dovey, Seth Giddings, Iain Grant, and Kieran Kelly. 2009. *New Media: A Critical Introduction*. 2 edition. Milton Park, Abingdon, Oxon ; New York, N.Y: Routledge.
Meyer, Michael. 2001. "Between Theory, Method, and Politics: Positioning of the Approaches to CDA." In *Methods of Critical Discourse Analysis*, by Ruth Wodak and Michael Meyer, 14–31. London: SAGE.
Miller, Claire Cain. 2015. "Algorithms and Bias: Q. and A. With Cynthia Dwork." *The New York Times*, August 10. http://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html.
Moretti, Marcus. 2014. "How a Simple Google Image Search Exposes Society's Most Harmful Stereotypes." *Mic*. December 31. http://mic.com/articles/107468/what-google-image-search-reveals-about-our-cultural-stereotypes.
NVivo Qualitative Data Analysis Software. 2014. QSR International Pty Ltd. Version 10.
"Our Secret Sauce." 2015. *Official Google Blog*. Accessed October 28. https://googleblog.blogspot.com/2008/02/our-secret-sauce.html.
Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. EBook. London: Penguin Books.
"Policies." 2015. *Google Inside Search*. Accessed October 22. https://www.google.com/insidesearch/howsearchworks/policies.html.
Reisigl, Martin, and Ruth Wodak. 2001. *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*. Routledge.
Rubinstein, Daniel, and Katrina Sluis. 2013. "The Digital Image in Photographic Culture." In *The Photographic Image in Digital Culture*, edited by Martin Lister, 2nd ed., 22–40. Routledge.
Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. http://www-personal.umich.edu.proxy.library.uu.nl/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.
"Ten Things We Know to Be True." 2015. *Google Company*. Accessed October 25. http://www.google.com/about/company/philosophy/.
Vaidhyanathan, Siva. 2011. *The Googlization of Everything.* Berkely: CA: University of California Press.
Winner, Langdon. 1980. "Do Artifacts Have Politics?" *Daedalus*, 121–36.

Wodak, Ruth. 2001. "What CDA Is about - a Summary of Its History, Important Concepts and Its Developments." In *Methods of Critical Discourse Analysis*, edited by Ruth Wodak and Michael Meyer, 1–13. Introducing Qualitative Methods. London: SAGE.

Wouter Van Noort. 2015. "Europese Commissie Klaagt Google Aan Voor Machtsmisbruik." *NRC*. April 15. http://www.nrc.nl/nieuws/2015/04/15/europese-commissie-klaagt-google-aan-voor-machtsmisbruik-2/.

Wurth, Kiene Brillenburg, and Ann Rigney, eds. 2009. *Het leven van teksten: Een inleiding tot de literatuurwetenschap*. Amsterdam: Amsterdam University Press.

Appendix A: List of codes used for analysis

List distilled from Baker and Ellece's *Key Terms in Discourse Analysis* (2011):

- Absence and silence
- Access
- Agency
    - Passivization
        - Passive agent deletion
        - Passivized social actors
        - transitivity
    - Objectification
    - categorization
- Argumentation strategies
    - topoi
    - Rhetoric
- Collocation
- Connotation
- Genericization
- Intensifying strategies & mitigating strategies
- Legitimation
- Modality
- Overwording
- Personalisation
- Personification
- Metaphors and similes
- Presuppositions