ROYAL HOLLOWAY, UNIVERSITY OF LONDON

INFORMATION SECURITY GROUP

CENTRE FOR DOCTORAL TRAINING IN CYBERSECURITY

Google and the Algorithm:
Side Effects of Search

Author:
Pip THORNTON

Supervisors:
Prof. Keith M. MARTIN
Prof. Peter ADEY

September 17, 2014

Abstract

This project examines the powerful effects search engines (specifically the monopoly holders Google) and their algorithms have on society, knowledge, security and language. Algorithms have agency, but conversely they are also shaped by the people who use them, categories which are complicated by the commercial nature of Google and the blurred definitions of customer/consumer and user. Further to this, the data available to the algorithms is unrepresentative of the society it claims to mirror, with hierarchies of knowledge and experience foregrounded at the expense of the less technically and politically adept. With an unreliable database of reference, search results may not be as 'innocent' as the scientific method might suggest. There has been much recent debate about the power of algorithmic data over and between a range of academic disciplines and from a newly aware post-Snowden public. I aim to add to this debate by proposing a new critical theory of Reconstruction as a tool to re-liberate both language and society from the constraints of the algorithm and work out how to recognise, manage and treat the side effects of Search.

Contents

List of Figures

Chapter 1

Introduction

This is a broad reaching topic, each part of which deserves expansion, but I realised fairly early on beginning my research that I had a firm argument I wanted to present which requires each part of this essay to bring that to a conclusion as practically and succinctly as possible. I hope to be able to expand on this argument in future work.

In 2000 Introna and Nissenbaum wrote a prescient paper on the politics of search engines (Introna and Nissenbaum, 2000) in which they tried to make sense of the then rapidly emerging phenomena of the search engine, stressing the crucial importance of maintaining some kind of control over its already visible power and influence. They anticipated the commodification of information which online search effected, and also its situatedness in and power over the market, and called for public oversight of search engine operation (see also Hinman, 2008) and transparency and disclosure of the algorithms. But what they did not anticipate was that one search engine would quickly gain a monopoly, not only of search, but of the market forces inscribed within it. In the last quarter of 2013, Google had a market share of 91% in the UK, 71% in the US, and 96% in Germany.[1] Only in China, where the search engine is less dependent on global advertising revenue, has a search engine (Baidu), been able to compete against the giants Google and Yahoo[2] and meanwhile 'to Google' has entered the vernacular, and Google apps, schemes and projects have expanded to mapping the world, digitising its literature, keeping its secrets, anticipating its questions, and even driving its cars. Google has become indispensable, and all of this with a lack of transparency so subtle that most people don't care or think to question its motives, accuracy or objectivity, let alone its political implications or the potential effects it is having on society (Tufekci 2014) , space (Graham 2005, Kitchin and Dodge 2011) , language (Fuller and Goffey 2009, 2012) and culture

---

[1] http://blogs.adobe.com/digitalmarketing/advertising/continued-growth-googles-search-ad-business-q2/

[2] Baidu is funded by advertising, but it is local in nature and, like the code on which the platform depends, is exclusively Mandarin

(Rose 2014). "How did so many people end up being that dependent on a single search engine?", asks Geert Lovink (Lovink 2009).

In recent years there has been a explosion of work around net neutrality, privacy, human rights, security, software sorting, the status and influence of search engines such as Google, and much of this work has concentrated specifically on how data and algorithms affect us - perhaps somewhat ironically because researchers have found the algorithm to be fundamentally unstudiable, not least because it is black-boxed technology which maybe even the Google coders and engineers no longer have control of (Morozov 2012) . Algorithms and their data are potentially geographically and socially biased and are also unstable - mobile, massive and mutable (Rose 2014), with all the security concerns that might entail, and not only can their product change from minute to minute, place to place, person to person in terms of various levels of personalisation, their workings can be altered or tweaked by Google for many reasons and purposes, for quality control, censorship or commercial gain, at any time, with no notification or justification. Embedded in and co-dependant on this privately owned and ostensibly unregulated technology are the flesh and blood users of Web 2.0 and beyond, with all the ethical, legal and moral issues which that might entail (Sandvig 2014, Feuz 2011).

What is clear is that the study of algorithms and specifically online search is methodologically challenging, with limited benefit in empirical research or attempts at reverse engineering because of the obfuscated and transient process of their existence. Ostensibly hidden from view to stop spammers and protect advertisers, the algorithms are left to do their business on their own - or so we are led to think. Just how much human input is there in an algorithm? How much tinkering is done behind the scenes for research, social engineering, commercial, censorial purposes, and how much human residue accidentally remains in these programs, codes and algorithms, and therefore potentially in their results?

Algorithms need to be studied as best and as critically (Thatcher &Dalton 2014) as we can, not just as technical, but as political issues (Introna &Nissenbaum 2000). If in the past there was a certain degree of apathy towards the purely commercial nature of potential search engine bias in industry and academia (Van Couvering 2010), the prominence of the search engine in social, cultural, ontological and epistemological ways, as well as a producer of opinion, has changed that and galvanised the academy into critical exploration.

So is Google really "holding the world by the thoughts" (Battelle 2005), or is

its power exaggerated? Some see the search engine as an insidious agent and supporter of capitalism (Jarrett 2014), or even as a potential replacement for democracy (Aaronson 2014), morality and justice (Goldstein 2014). Introna (2013) sees it less of specific algorithm problem but as part of wider network, and others see Google as just another evolutionary and unstable step in information processing (Darnton 2009).

This project will therefore concentrate on what we can see and can study, which is what we put into a search engine and what comes out, and the effects on society, structures of language and knowledge. The concepts of power and security permeate at every stage, and will form a theoretical framework for discussion. I use the term security deliberately ambiguously to highlight the tensions between good and bad security, but at the same time try to avoid such rigid structure - taking into account the subjectivities, tensions and hierarchies of power, knowledge and necessity which compete in the now near ubiquitous world of online search.

For the purposes of this project I have split Search into three main sections which are loosely based around what computer scientists Friedman and Nissenbaum (1996) identified what they believed to be three points at which computer systems could be biased: "three categories of bias in computer systems have been developed: pre-existing, technical, and emergent. Pre-existing bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use." I want to apply these categories to the issue of search. The pre-existing category is the context (Production of Context), the technical category is the algorithm itself (what the algorithm does, the technical - although black boxed so difficult), and the emergent category is the result - what happens at the end - the 'product', not only the search result and their impact, but statistics, autopredict, language knowledge, privacy. Woven through these categories are issues of security, power, and knowledge, from Fairclough's 'soft power' of language (2001),to the threat of harsher, more disciplinary Foucauldian biopower. It is important to differentiate between the database and the algorithm (Gillespie 2014), which is also reflected by van Couvering (2010), who separates the operation of a search engine into three categories, all of which can be biased: the index (by categorisation bias, seed lists and index cleaning), the algorithm (by its creators, ranking systems and links, relevance etc.) and display results (largely as result of advertising, bigger results, colour, font etc.).

Chapter 2

The Production of Context

> When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure. There was no personal or world problem whose eloquent solution did not exist in some hexagon... As was natural, this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible seemed almost intolerable. (Borges, 1964)

It might be a cliche and seem blindingly obvious, but in terms of search, you can only get out what you put in. If something does not exist on the web, then it is not going to be found (or auto-suggested) by a search engine. This raises a few points, and especially what it is 'to exist' on the web, which means not only data being part of the web database, but also its capacity to 'exist' to a search engine, because if not visible (or readable) to a search engine information, information on the Web might as well not exist (as is the crux of the current spate of 'Right to be Forgotten' applications being made to Google). This chapter will examine what and whose data is and is not on the Web and why, as well as the reasons data on the Web might be less visible to search engines. Without a database of information, algorithms would have nothing to work on (Gillespie 2014), so being aware of what they can or can't work on is vitally important in assessing what they produce. "Make no mistake", warn Introna and Nissenbaum, "these are political issues. What those who seek information on the Web can find will determine what the Web consists of - for them." (Introna and Nissenbaum 2000) The chapter title is deliberately Lefevbrian[1]: The production of the contextual spaces of the web is as socially produced as physical space, and as such reflects an economically and demographically unequal society that, as I will argue, through search, reproduces

---

[1]French theorist Henri Lefevbre's The Production of Space

that inequality algorithmically, recursively and ad infinitum.

2.1 Big Data

Already suffering from "data obesity" (Lovink 2008), the amount of data held on the Web is 'big' and growing every second. The quality of that data, however, is decreasing exponentially(Feuz 2011), presenting a challenge to search engines to which they have responded with new tactics such as personalised search, a development which in effect leads to the possibility of big data being used to (or incidentally) for surveilling and engineering the public. (Tufekci 2014)

Big Data has also posed interesting questions in academia, forcing "a radical shift in how we think about research... a profound change at the levels of epistemology and ethics" (boyd and Crawford 2012) . The ability to collect, analyse and extract meaning from huge swathes of empirical data on a scale impossible before the Web has been seen by some as the 'end of theory' - the dawn of an era where knowledge emerges from scientific rather than human method (Anderson 2008).But there are several reasons to be careful of over-emphasising the importance of using big data as a sociological tool (boyd 2012), not least because despite having the air of legitimacy due to its vastness and apparent inclusiveness, the data (however big) still runs the risk of being unrepresentative, as I will be arguing in the next section, and tainted by the processes which generate and sort it (Gillespie 2014). Far from being a democratic rhizomatic Web in the Deleuzean sense, "users of the Internet are taken through tracts of knowledge rather than connecting through random intersecting points" (Hess 2008), and even big data will struggle to even out those tracts.

2.2 A Representative Database?

> It is impossible to organize the world's information without an operating model of the world. (Stalder &Mayer 2009)

2.2.1 Geopolitical Data

As Crawford reminds us, "Data and data sets are not objective; they are creations of human design" (Crawford 2013a), and as such are just as situated geographically and politically as their designers. Big data will always provide an "oligoptic... partial and selective" view of the world (Kitchin 2006) because not every human

has the same access, digital agency or data footprint, which begs the question how representative is the data that exists on the Web? Or as Neyland puts it: "Who and what is included or excluded, on what terms and to what end?" Although technically 'anyone' can add to the Web database, there are whole communities who, whether due to a lack of access to technology or utilities, or a lack of technological ability or education, have never done so, and similarly will have no input in how it is controlled, processed or analysed. "Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities", writes Crawford (2013a), and uses as an example analysis of data from Twitter during Hurricane Sandy when the volume of Tweets about the disaster came from Manhattan, not the affected area, which reflects not only the hole in the dataset due to the emergency, but also a social bias towards people most likely to engage on certain platforms. So, "while massive datasets may feel very abstract, they are intricately linked to physical place and human culture", and may cut out sections of society by class, economics, geographical location (places with no signal, for example). The map at Fig.2.1 gives some idea of the concentration of Web access in Europe and North America, although the somewhat surprising relative darkness over China has its own geopolitical significance in the blocking of the IP 'pinging' software used to create it.

As I will discuss in the next section, the way data is represented, organised and used on the Web is also inherently political, with the algorithms that sort and process, magnify and duplicate data adding another layer of weakness to their integrity. For all its promise of accessibility, democracy and public empowerment, the Web has, say Introna and Nissenbaum (2000), gone down exactly the same route as previous media revolutions like TV and radio, which also promised these things, but ended up with power in the hands of a few dominant people and institutions. In terms of data analysis too, the inequalities of Web access and agency have led to a "data-analytic environment that favours the powerful, data-rich incumbents, and the technologically adept". (Tufekci, 2014) In addition, some scholars argue that the physical location and structure of search engine companies and servers which hold the data plays a part in the potential bias of search engine systems. (Vaughan & Thelwall 2004, Van Couvering 2010)
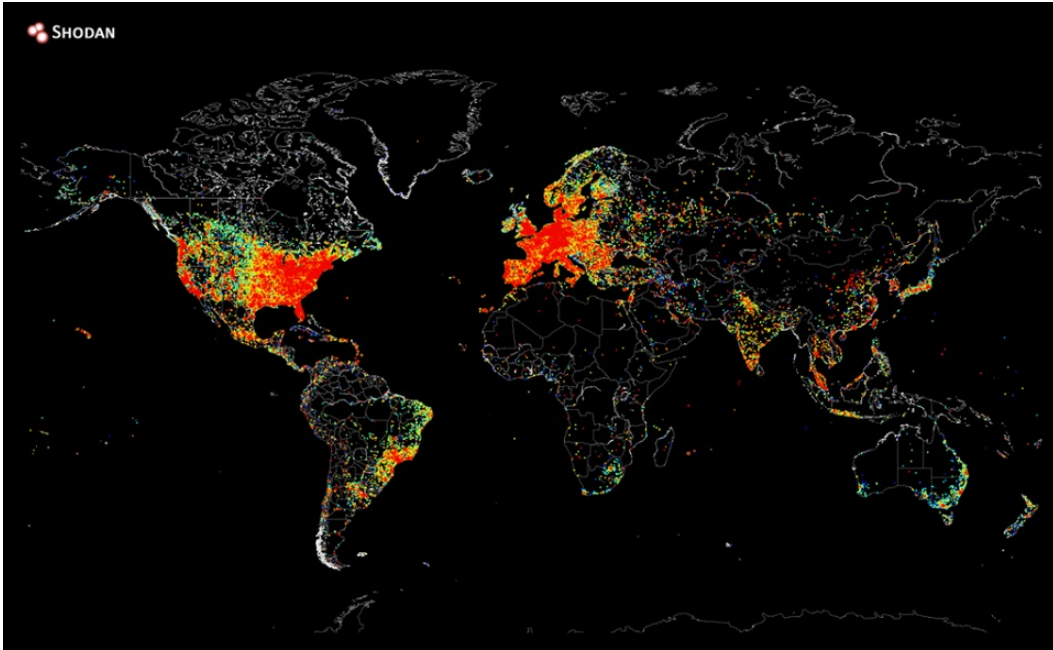
Figure 2.1: Map of internet-connected devices August 2014, by internet cartographer John Matherly http://www.shodanhq.com/

2.2.2 Textual Data

As I said before, the pool of searchable data on the Web has been put there by human choice. In 2004 Google launched Google Book Search, a scheme (now known as Google Books) which aims to digitise every book and article in the archive. The scheme has been heavily criticised for the unsystematic and haphazard way it was executed, Google employees simply scanning every single book in the libraries geographically closest to the Google HQ, regardless of bibliographic significance (see Darnton 2009) and also necessarily predominantly texts written in English and specifically North American in provenance and influence. Particular opposition was raised by the president of the French Bibliotheque National JeanNoel Jeanneney:

> We are faced with several possible dangers: with respect to works of various cultural heritages that have fallen into the public domain, the list of priorities will likely weigh in favor of Anglo-Saxon culture. With respect to works still under copyright, of which on excerpts, or 'snippets' will be offered for the time being, the weight of American publishers may be overwhelming. As for journals and books disseminating on-going

11

research, the dominance of work from the United States may become
even greater than it is today. (Jeanneney 2008, 6)

The importance of which, how many, and in what order words are added to the searchable database will be explored later, but it is clear to see the power that Google wields in terms of constructing it linguistically as well as culturally. As Jeanneney also points out, a digitised library of predominantly English (North American) texts brings with it the weight of content too, with cultural, administrative and societal traits foregrounded through volume alone at the expense of other languages, cultures and systems, which are left forsaken at the "gondola end" of the digitised database. There is also a hierarchy of index updating on the Web, with search engine crawlers revisiting established blogs and news sites with a high turn over of content much more frequently than others, which makes sense in terms of keeping 'up to date', but also means that the dominant content is constantly reproduced, re-indexed and re-cached, adding to the pool of contextual data from which search results can be drawn.

2.2.3   Factual Data

Mark Graham has recently pointed out similar content bias in Wikipedia (Graham 2014), a platform which by nature of its user interaction and editing, is often held up as a success of the democratic Web. But an analysis of geotagging shows that the 'facts' presented by Wikipedia are constructed by and of the world's dominant languages and places. "Knowledge created in the developed world appears to be growing at the expense of viewpoints coming from developing countries", writes Graham, with 84% of geotagged articles being about Europe and North America, and more articles about the Middle East in English than in Arabic languages. It is easy to see how this might also effect Google Search, especially as Google now use Wikipedia as a source for its 'intelligent' Knowledge Graph feature which pulls facts and images together into a handy box displayed at the to of your search results. Added to this, by virtue of its link structure and reputation (which will be discussed later), Wikipedia has been found to be on Page One of search results in 99% of searches.[2] Therefore as Graham suggests, "Wikipedia might not just be reflecting the world, but also reproducing new, uneven, geographies of information."

---

[2]http://searchenginewatch.com/article/2152194/Wikipedia-Appears-on-Page-1-of-Google-for-99-of-Searches-Study

(Graham 2014) The tendency for search results to favour collected information or analysis sites is, according to Reilly, also at the expense of original (potentially controversial or partisan) source material, which can lead to minority non-state groups or actors being made less visible, (Reilly 2008) and results being dominated by bland links and under the influence of groups with more commercial or political "organizational clout" and "off-web prominence"(Gerhart 2004). Thus, because of the ranking algorithms, the actual objects of the search become invisible, replaced by representations of themselves, whether an opinion, a encyclopedia type entry or an advert (Mager 2012), all the time the search engine is further abstracting the user from the object of their search.

2.2.4 Personalisation

I will be discussing the effects and implications of personalised search later on, but it is important here to point out that the search results the user receives can be affected before you have even typed anything into the search field. Google reports that there are over 200 different signals which determine the results of searches. Factors such as your device, your location, domain (google.co.uk, google.com etc.), search history (of the machine, so not necessarily your own search history), and the time of day may put the user in a certain category and will therefore affect the information retrieved, as will the language and medium used. Using Google Scholar, for example might suggest a certain type of person or class, and as we shall see, all these factors are used to maximise revenues by advertisers and by Google. In these circumstances, what you type is less relevant that the situated type of audience you are perceived to be. This can be seen as a "shift from keywords to audiences".[3]

2.3 Organising, Indexing and Categorising

2.3.1 Bias in Indices

According to Gillespie, categorization is a powerful semantic and political intervention" (Gillespie 2014), and identifying, defining and cataloguing bodies has long been a form of power over a population, controlling space (see Graham on Software Sorting), international movement (Amoore 2006) and societal roles and expectations. In

---

[3]http://blogs.adobe.com/digitalmarketing/advertising/continued-growth-googles-search-ad-business-q2/

terms of information, the restrictive and hierarchical consequences of categorising are also far-reaching, (see also Becker 2009) and shaped by the intent and beliefs of the dominant indexer, as highlighted by the example of Melvil Dewey (of Dewey-Decimal fame) whose original catalogue put all "religions other then Christianity" into a single category. (Stalder & Mayer 2009, 98)

2.3.2 Segmentation of the Web

Google's ambition has always been to "organize the world's information and make it universally accessible and useful"[4], but as we have already seen, for reasons of geopolitical, social and technical inequalities, the concepts both of 'the world's information' and 'universally accessible' are both problematic, with access to the Web for information input and retrieval hampered by a 'digital divide" (see Fuchs & Horak on Africa and Ragnedda & Muschert on global divides). which, if further distorted by segmentation "may merely mirror societal institutions and their baggage of asymmetrical power structures, privilege, special interests, and so forth." (Introna and Nissenbaum 2000, 61)

2.3.3 Invisible Barbed Wire

However, Google attempts this improbable task by sending crawlers to trawl and digitise everything on the Web from searches to social networks, and this digitised information is then assessed and algorithmically 'marked-up' and placed in a category so it is more readily retrievable should that category be searched on. It could be something as simple as a web page being categorised as a blog, chatroom or news site, or be based on what the algorithms think are the salient or 'key' subjects to tag the page with. How accurate these digital assumptions are is hard to assess and might only become apparent through the appearance of a poor or irrelevant search result. But when it's personal data being categorised and tagged, for example in the case of personalised search, such categorisation can have substantial effects in "losing control in defining who we are online... we are losing ownership over the meaning of the categories that constitute our identities", warns Cheney-Lippold, we are 'free' but "constantly conditioned" (Cheney-Lippold 2011, 178). Some highly sensitive personal data, relating to gender, sexuality or class, for example might also be more readily computed if it fits into the most clear-cut or 'normative' groups,

---

[4]https://www.google.com/about/company/

reproducing the dominance of fixed categories at the expense of minorities or a less rigid structure of social categorisation (Feuz et al. 2011, 1). This categorisation, although a means of control in itself, a kind of "panoptic sorting" (Gandy 1993, Lyon 2003, Graham) , can be all but invisible to the data subject and consequently difficult to resist. As Morozov notes:

> The invisible barbed wire of big data limits our lives to a space that might look quiet and enticing enough but is not of our own choosing and that we cannot rebuild or expand. The worst part is that we do not see it as such. Because we believe that we are free to go anywhere, the barbed wire remains invisible. (Morozov 2012)

While we may not be able to resist the categorisation of our data, it is possible to make the choice to block web crawlers in order to remain off the index, and of course this process can, as I will expand on in the next chapter, be implemented by Google itself in some cases involving litigation, for example in 'right to be forgotten' cases.

2.3.4   The Right to be Forgotten

When the European Court of Justice ruled against Google in the case of a Spanish man who wanted historic information about him removed from the searchable index, much of the resulting coverage used the library analogy to explain how the Google system works. Removing the link to a piece of information on the database does not delete the page bearing that information, but just like removing a physical card index page from a (non-digitised) library doesn't destroy the book to which it refers, it just means you cannot now find out where it is. This is an incredibly strong hold to have on access to information and knowledge, especially as it is impossible for the user of a search engine to know if links have been removed, as although Google notify the page owner (as happened recently with Robert Peston)[5], and have now taken to adding a caveat to all results pages which relate to a name. Gillespie refers to the "archive's jussive force", which "operates through being invisibly exclusionary (Gillespie 2014), i.e. the archive presents itself as being the set of every possible statement or thing, but is actually just what Foucault calls "the law of what can be said, the system that governs the appearance of statements as unique eventsit

---

[5]http://www.bbc.co.uk/news/business-28130581

is the system of its enunciability," (Foucault 2012(1969)) and therefore not only reproduces information, but is part of a productive discourse itself, which I have been arguing is precisely the case with the modern day archive of searchable data.

I will be discussing the security issues - both technical and societal, which come with online search in the next chapter, but will point out here how similar in structure and function the search engine is to an intelligence agency such as the NSA or the Security Service. The former head of MI5, Stella Rimington always stressed that the relatively unexciting jobs of filing and indexing were actually the the key to success: "A Security Service lives by its records", she wrote in her autobiography (Rimington 2001, 94). Of course Google has been accused of (but strongly denied) collusion with the intelligence agencies in the aftermath of the Snowden revelations, but whether they did allow backdoors or share data or not, the parallels between their systems, and even the skill sets of their workforces[6], are obvious, leading some to fear a day when "algorithms could trap us in a world where advertisers and government agencies couple behavioural data with computer formulae to predict and manipulate what we do or buy next[7].

2.4 Hierarchies of Knowledge

So if Google Search holds the index of a huge library, there is a considerable gap left by the absence of a librarian. In a library there is implicit trust that the person you ask for help - the person who has probably also categorised and indexed each book on the shelves, is qualified to do so, and is knowledgeable, objective and trustworthy enough to guide you to the 'best' information. Having already discussed access to information then, this section will look at how the digital search for data deals with questions of quality, judgement and hierarchies of knowledge.

2.4.1 Quality of Content

"With the rise of Internet search engines", writes Geert Lovink, "it is no longer possible to distinguish between patrician insights and plebeian gossip" (Lovink 2009, 45). With no gatekeeper to point users in the direction of the most accurate or reliable information, users have to rely on the algorithms to do it for them, which

---

[6]http://www.reuters.com/article/2014/08/07/us-tech-nsa-idUSKBN0G72BG20140807
[7]http://www.pcadvisor.co.uk/features/internet/3304956/how-google-facebook-and-amazon-run-the-internet/

is problematic as the algorithms - unlike the wisdom of the librarian - are based on popularity not truth (see PageRank section in next chapter). There are, as Gillespie suggests, "palpable but opaque undercurrents that move quietly beneath knowledge when it is managed by algorithms" (Gillespie 2014), undercurrents sometimes of dubious origin which may lead to a "decentring and deprofessionalisation of knowledge and professional expertise" (Jenks 2002, 74). According to Hinman "search engines have replaced scientific and scholarly legitimation with a digital version of the vox populi" (Hinman 2008, 67).

2.4.2 Knowledge Control

Lovink argues that in the era of Web 2.0, where information and knowledge is generated interactively by the public, "the scientific establishment has lost control over one of its key research projects the design and ownership of computer networks" (Lovink 2008, 1). Media and market forces have overtaken mathematics, and Google has grown too quickly for it to be controlled or mediated or understood by the intellectual elites who would traditionally moderate new things (Lovink 2008, 2), its distributed structure unsuited to a centralised index or catalogue (Becker 2009). Knowledge is therefore being produced in a different way, unregulated by the traditional scientific, academic gatekeepers, but now subject to the standards, influence and motives of a private company. By using Google as both library and librarian we have "systematically outsourced our sense of judgement to this one company, we've let this company decide for us what's important and what's true for a large number of questions in our lives." (from short interview with Vaidhyanathan by Network Cultures 2013)[8]

---

[8]http://vimeo.com/82099408

Chapter 3

The Illusion of Algorithmic Objectivity

The title of this chapter is borrowed from a paper given at the Governing Algorithms conference in New York last year by Tarleton Gillespie, which challenged both the assumption, and the acceptance that anything an algorithm does is automatically objective precisely because it is based on mathematics. This (somewhat ironic) 'faith' in technology and science over more theoretical or metaphysical system has a long cultural heritage which today is in part perpetuated by the common perception that it is all 'too hard' for most people to understand and therefore best left to the 'experts'. In terms of algorithmic objectivity, there are several reasons why we need to look beyond the numbers and avoid a 'simplistic technological determinism' which sees "algorithms as abstract, technical achievements, but must unpack the warm human and institutional choices that lie behind these cold mechanisms"(Gillespie 2014). Louise Amoore points out that even Alan Turing recognised that mathematics is a mixture of human intuition and machine ingenuity, arguing that maths is inherently political by nature (Amoore 2014, 6-7).

3.1   Subjective Data

As I have argued in the first section, we should look critically at hierarchies, inclusivities and lacunae of the data available to search engines, but also bear in mind the subjectivities, motives, backgrounds, affiliations of the person, or group of people who create, deploy and continue to control the algorithms which seek to sort that information. It is, as Gillespie suggests, a carefully crafted fiction that algorithms are neutral by their technical nature, yet it is a fiction strongly defended and maintained especially in the face of controversy (Gillespie 2014), as we shall see later (for example allegations of a racist/sexist Google).

A reliance on the technological purity of the algorithm brings with it assurances

that anything the algorithm produces will be free from influence or bias, will be fair, accurate, objective, neutral, and dependable, all backed up with it the authority of science. When we translate this to the potential affects this has on the users of search, such reliance creates confirmation bias, ontological securities, reifies logics and creates new knowledges, which at the same time galvanising the search engine (and its brand) in a reciprocal loop of trust and loyalty.

3.1.1 An Entanglement with Practise

Grimmelmann claims that "search engines pride themselves on being automated (Grimmelmann 2008, 950), but I think it goes further than that. The perception of the 'hands-off' nature of search engines is big business, and as such is carefully constructed. There is significant commercial value in being the 'fairest', 'most democratic' etc. search engine and being able to return better, more 'relevant', results. But this illusion of objectivity creates a tension or paradox between the actual algorithmic system and the appearance of what it does which Gillespie refers to as an "entanglement with practise" (Gillespie 2014)

3.1.2 Coders

Code and computer language are not objective purely because of their technical nature. As with literature, there is always a reason why it is written, be it for pleasure, for profit, for education or as some kind of statement, political, legal, subversive or otherwise. It should therefore always be treated as situated and embodied, and its interactions with the data it processes should also not be dismissed as free from their residual influence. And when the data comes out of the other end of a search engine, it still has to be analysed, interpreted and presented to the user. There are many opportunities and pressures for outside bias and influence in this process, just as there are when the algorithm is designed, written, tested and deployed. "Making sense of data is always framed", writes Rob Kitchin, "even if the process is automated, the algorithms used to process the data are imbued with particular values and contextualized within a particular scientific approach" (Kitchin 2014). Data will always bear the imprint of the inputter, analyst or interpreter of the data and consequently the residue of their geographical, habitual and social status and skills. Mackenzie sums this up nicely:

> Code can be read as permeated by all the forms of contestation, feeling,

identification, intensity, contextualizations and decontextualizations, sig-
nification, power relations, imaginings and embodiments that comprise
any cultural object (Mackenzie 2006, 5).

The context of coding is really important here, with market forces driving de-
mand and the search for better, more profitable algorithms. Gillespie highlights the
growing politicisation of the corporate entities behind search and data analytical
algorithms. Those who traffic in user datahave a stronger voice because of it, in
both the marketplace and in the halls of legislative power, and are increasingly in-
volving themselves in political debates about consumer safeguards and digital rights
(Gillespie 2014).

3.2   PageRank

Google's search algorithm is called PageRank, and was developed by founders Sergey
Brin and Larry Page as post-doc students at Stanford University in 1996 and first ap-
peared as an academic paper entitled *The anatomy of a large-scale hypertextual Web
search engine*. Based on the old-fashioned method of academic citations, PageRank
works on the basis of links between sites. A site with many inbound and outbound
links to other sites is deemed to be 'more important' than others. Added into the
mix is an algorithmic assessment of the quality of those links, which according to
Kleinberg can carry with them hierarchies of power, latent human judgement and
authority. (Kleinberg 1999) Generating false links to a site by spamming will there-
fore not necessarily help to raise the ranking of a site, although plenty of link-farms
and black-hat search optimisation organisations (SEOs) still try this method with
varying success.

PageRank still forms the basis of the Google search algorithm, but has had sev-
eral major upgrades over the years and is indeed updated (or tweaked) on an almost
daily basis, one of the reasons empirical study is so difficult. Many of the tweaks
made by Google are to control spam or what they judge to be underhand advertising
or rank boosting techniques, and can have devastating financial effects on legitimate
businesses who suddenly find themselves inexplicably at the bottom of search rank-
ings. One of the most recent major upgrades was the Hummingbird update which
took effect in August 2013, and aims to create a more semantic approach to search,
concentrating less on keywords, and more on trying to find sense from the string
of words entered. Hummingbird can be frustrating to use as a researcher, as it has

a tendency to second-guess what the user is looking for with frequent assumptions that the user 'meant something else', or misspelled a word. There are workarounds, including inverted comma and Verbatim searches, but these can be little known revert to default after each search. The Hummingbird algorithm dramatically changes the way results are generated, as I will show in the next chapter.

3.2.1   Shareholder Democracy

One result of Kleinberg's theory of *authoritative sources in a hyperlinked environment*, is the possibility that the link system of ranking carries with it a democratic representation of society, or a "shareholder democracy":

> One link is not one vote, but it has influence proportional to the relative power (in terms of popularity) of the voter. Because blocks of common interests, or social factions, can affect the results of a search to a degree depending on their relative weight in the network, the results of the algorithmic calculation by a search engine come to reflect political struggles in society. (Finkelstein 2008)

Others such as Scott Aaronson, have developed this idea further. Drawing on an idea from a pre-Google link-ranking system developed (and subsequently abandoned) by his professor Jon Kleinberg (as above) at Cornell University, Aaronson discusses whether the idea of authoritative link structure of search engines could be used to solve classic metaphysical and philosophical problems such as the measure of morality and justice, as well as creating a kind of 'Eigendemocracy'. (Aaronson 2014) He also cites a recent book by Rebecca Goldstein called *Plato at the Googleplex* (2014), which imagines what an internet-savvy Plato would make of the crowdsourcing of knowledge via online search.

But such philosophical and playful musings are based on the 'idea' of the egalitarian algorithm and not its practice. As I have shown, PageRank is marketed as "fundamentally democratic computational logic" from the start (Gillespie 2014), but it is inherently hierarchical and unstable, and is constantly tweaked from within to reorder results. It is no basis for Socratic dialectics.

3.2.2 What is relevant?

Who decides what is a 'relevant' result? It is a subjective quantity which it is as problematic to measure as it is to define. Clicking on a search result, for example, does not guarantee that it was the most relevant or 'best' result there, or even that it was what the user was looking for. Some empirical work is done on "Feedback Loops from User to Algorithm' (Granka 2010), but as Gillespie suggests, satisfaction does not necessarily equal relevance (Gillespie 2014), nor does success equal quality (Van Couvering 2010).

3.3 Manipulation of Search Engine Results Pages (SERPS)

So by way of summary so far, there seems does indeed seem to be an illusion of algorithmic objectivity, yet as users we still appear to put faith in algorithms which a) may be biased and based on biased datasets, and b) may be manipulated by humans anyway, so are not objective in either way yet still have the power of people's trust and also (as I will go on to show) have a substantial effect on language and culture, society and knowledge. Not only that, but as the following sections will show, the search process is a commercial, private, corporate, money making, unaccountable, black boxed system which holds personal data on a scale bigger and less accountable than state intelligence agencies, yet the data is only 'secured' on a customer-provider system of trust which itself is problematised as the 'customers' are not the actual search users, but the advertising agencies whose commercial success funds the Google empire.

3.3.1 Categorisation

Google Trends and Google Analytics are popular and useful tools for companies and researchers alike, but it is important to realise their shortcomings. Just like indexing forces strict tags of meaning onto data, there are also potentially restrictive undercurrents beneath the concept of 'trending' searches and data. As Twitter has shown, tagging words and phrases makes them easier to search and consequently more prominent, but the popularity of 'trending' subjects is heavily controlled (in the case of Twitter), and severely delayed and diluted in the case of Google. Google Trends works on the same kind of semantic guesswork as the new search algorithm, meaning that what is actually searched for can easily be misreported, as in the case

of searches for Ferguson in August 2014, whereby the analytics suggested that many people who searched for the riot-stuck Missouri town also searched for a company called Ferguson Builders. There is also evidence that 'trending' algorithms reward spikes in traffic, which can lead to the concealment of slower-burning topics. (Tufekci 2014)

3.3.2 Hands-off Results?

As Gillespie notes, "algorithms can be easily, instantly, radically and invisibly changed(2014), so while we can never really know just how much behind the scenes tweaking goes on, there are many examples of Google overtly doctoring, censoring or judging search results. The Google website states that it does apply a narrow set of removal policies for pornography, violence, hate speech, and terms that are frequently used to find content that infringes copyright (Google Autocomplete Support). What is missing from this statement of course are definitions of terms such as 'hate speech' and 'pornography' and the scales of their unacceptability. Who decides these questions, the answers to which control what information the world is allowed to see? The results which do get through the self-imposed and unaccountable filters are presumably "scrubbed of the obscene, the objectionable, or the politically contentious, (Gillespie forthcoming) but which laws or guidelines dictate this, and which people, groups, companies or countries do they claim to represent? The famous example of overtly "hands-on' approach to result manipulation came in when Google agreed to be censored by the Chinese government, eventually pulling out of the situation in 2010, but are still adamant that users can "trust our objectivity" in less controversial domains, despite making and publishing judgements on some of the less savoury results which its algorithms produce, for example the Jew Watch incident in 2004, when Google left the anti-Semitic site at the top of the results, but added a caveat to publicly distance themselves from the sentiment.

Google also has an admittedly biased policy on advertising: "We dont try to put our sense of ethics into the search results", Sergey Brin was quoted as saying in 2004, "but we do when it comes to advertising".(Sheff 2004) Diaz lists numerous examples of Google's advertising bias between products which are mentioned in the Playboy interview: beer not wine, porn not guns, and also bias towards sponsors with controversial or non-Google friendly links (Diaz 2008). Google promises its

users that "no short-term gain could ever justify breaching their trust"[1], but as the next section will show, short-term gain is Google's lifeblood.

3.4 The Economics of Search

Despite its altruistic claims, Google is a business, not - despite the similarities - a public library or a state intelligence agency. As Feuz notes, "over time, it transferred itself into an advertising company, producing not search results, but audiences as its primary commodity" (Feuz 2011, see also Vaidhyanathan 2011, 3) This duality of function is described at a Second Index by Stalder and Mayer (2009), the first index is the supposed free association web where the supposed users/customers can get their search information, and the second, which is where the real customers (the advertising agencies) can get their ready categorised and key worded, classified (in class sense too) data so they can target the search users (whose data they have harvested) for money. Rieder and Sire identify Google Search as a "three-sided market" for this reason (Rieder and Sire 2014, 200), and it is easy to see why Google attracts some fairly strong Marxist critique (Jarrett 2014, Pasquinelli 2009, Fuchs 2011).

3.4.1 Marxist Critique

Not only is "insight into the workings of information algorithms ... a form of power" (Gillespie 2014), but it is also a source of money, and the principles of capitalism are embedded in the workings of search engines. (Mager 2012 and van Couvering 2010)

> Google is a meta-exploiter of all user-generated content producers. Users employ Google services and thereby conduct unpaid productive surplus-value generating labour. They engage in different unpaid work activities (searching, e-mailing, creating documents, blogging, reading blogs, uploading videos or images, watching videos or images, etc.) (Fuchs 2011)

So Google users are both consuming search and producing new data/profit. We have become what Alvin Toffler termed 'prosumers'. (Toffler 1981), doubly commodified first as data producers, and secondly our very consciousness is commodified

---

[1]https://www.google.co.uk/about/company/philosophy/

by permanent exposure to advertising our data has generated, (Fuchs 2011). Google is charging us "cognitive rent". (Pasquinelli 2009) What makes this unpaid labour even more exploitative is that many search users are so far alienated from the value they are adding (Jarrett 2014), as I will show with the Google reCAPTCHA program in the next chapter.

3.4.2 Semantic Capitalism

I will be discussing the commercial and epistemological power of semantics in the next chapter, but it is helpful to mention here that Google Search survives on the commodification of language, or as Feuz calls it "semantic capitalism... using control over words and symbols as a means of expanding capitalistic logic" (Feuz 2011). The keywords which (re)direct the search user from their search results are valuable commodities (Lee 2011), as are the "warehoused tracts of lexical space" squatted by opportunistic domain names.(Fuller and Goffey 2009) "Any word of any language has its price, fluctuating according to the laws of the market". (Bruno 2002) Even trade-marked words are not safe from the clutches of the market, as US company American Blinds discovered when they protested about the auctioning of their brand name.[2]

3.5 Manipulating Results

As mentioned briefly before, although Google do block suspicious behaviour, people so still try to game the PageRank link system for humorous, commercial, but even perhaps more sinister reasons.

3.5.1 Googlebombing

Grimmelmann (2008) describes the birth of Googlebombing and how search algorithms can be manipulated. It can be used in jokey way the founders of googlebombing managed to make their friends names return as top results on the searches 'Internet Rockstar' and 'Talentless Hack', but more seriously in 2003 democrat users managed to link George W Bushs biography page with the phrase 'miserable failure, and this is all done through links the actual phrase doesnt appear on the page in question. Grimmelmann calls this a "significant new form of politicking because

---

[2]http://www.reuters.com/article/2007/09/04/us-google-trademark-idUSN0336124420070904
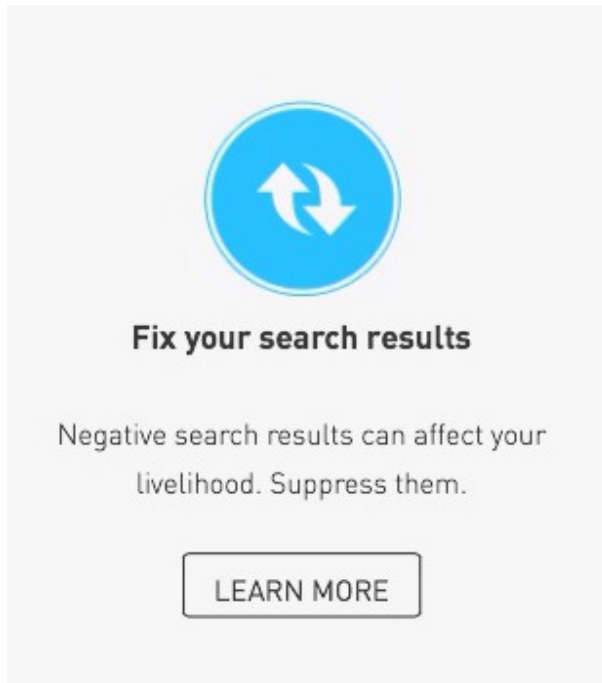
Figure 3.1: Screen Shot from Reputation Management Site http://uk.reputation.com/

people rely so much on the 'truth of Google. "Land a bomb like this and you can convince the world that Google agrees with your position (Grimmelmann 2008). A successful Googlebomb doesnt just reflect the consensus of web users; it can help construct that consensus', thus showing the potential agency of the algorithm and the ease by which it is manipulated.

3.6 Search Engine Optimisation (SEO)

SEO is a whole industry by itself, using either black hat or white hat methods to manipulate search rankings for a variety of reasons. SEO companies operate within the law (white hat) to optimise search engine rankings by analysis and promotion of keywords, content and HTML tagging (on page SEO), and by off page SEO which involves freshness of links and promotion of mentions etc. It is primarily used in a commercial capacity, but is also used for ORM (Online Reputation Management) to "suppress" information (see Fig.3.1.

Rohle notes how SEO companies are subject to a system of punishment and

reward (see Rohle 2009), which is in a Foucauldian sense a return to an overtly disciplinary form of security which seems in conflict with the freedoms the Web was meant to be heralding. (see also Klauser et al. on the disciplinary and potentially and regressive tendencies in the coding of utilities).

Chapter 4

The Agency of the Algorithm

"Algorithms have consequences" is a phrase used very recently in an article by Zeynep Tufekci (2014b) called What Happens to #Ferguson Affects Ferguson: Net Neutrality, Algorithmic Filtering and Ferguson. Tufekci was referring to how unequally the trend and newsfeed functions on media platforms treat different sections of society geographically and temporally. Now that users can interact with 'Web 2.0', the software no longer just mediates but now constitutes society (Beer 2009), and the algorithm is not just a processing tool but becomes part of its own result. But if algorithms are unequal in application but also generate, not just process, bias, there is then the question of whether people are aware of this potential algorithmic bias, 'curation' or manipulation (Hamilton et al. 2014, Sandvig 2014), and whether this matters?

Also potentially muting any power the algorithm may have in and of itself is its place as just one part of a network of affect (see Bruno Latour Actor Network Theory), any agency it might have being defined and qualified by the other actors in the network (for example the user, the programmer, the physical interfaces such as the screen, computer or modem (Galloway 2012). As Matthew Jones warns, "there's a grave danger in attributing too much agency to algorithms at the expense of the conditions that enable their performative effects." (Jones on Introna paper 2013 see also Barocas, Hood and Ziewitz 2013). But the networks, nodes and blackboxes of the Web are not themselves clean of human residue and influence and take the user though "tracts of knowledge", says Hess, "rather than connecting through random intersecting points of the rhizome. In short, the rhizome has grown trees." Algorithms can also be seen as a "communication technology (Gillespie 2014), i.e. they actually have agency in communicating and spreading knowledge.

Poon takes it further than this, seeing algorithms as the modern day facilitator of logistical progression, the movement of information as commodity being just as much of a 'revolution' as the movement of goods in the 19th Century:

> Algorithms do for information systems what canals did for mercantilism, or mechanics did for industrial production: the algorithm is an instrument of consistent replication that brings the spirit of industrial consistency to bureaucracy and information management" (Poon 2013)

But if we see the algorithm as a means of reproduction, what does this mean for creativity, or art? It brings to mind the classic Benjamin question of what happens to Art in the Age of Mechanical Reproduction? What then, happens to Art in the age of algorithmic reproduction? These are some questions addressed by Gillian Rose (2014) in her recent talk at the Royal Geographic Society. Other studies of search algorithms concentrate on the effects they may be having on our mental capacities. In providing information at the touch of a button what is the 'Google Effect' doing to our memory? (Sparrow 2011), and our unique "capacities of creative and thoughtful reflection" (Carr 2008), reading and writing (Stefan Weber on the Google Copy Paste syndrome 2007).

Totaro and Ninno explore what happens when algorithms go beyond ordering numbers and begin ordering language. Not only does this produce new knowledges, but it also suggests an actual action - an agency, or function beyond the production of knowledge.

> Once algorithms are applied to objects that apparently are not numbers, they take on a strongly practical connotation, going beyond the sphere of knowledge to invade the sphere of action. Thus, the logic of numerical functions is present in society not only as a cognitive language (i.e. as mathematics), but also as a performative one. (Totaro & Ninno 2014)

They cite manufacturing processes, provision of services, and clicks, but this must also apply to search engines, and means that the ''ogic' of numbers has a performative effect on us: "The logic of numeric functions enters the practical world, often unseen, and firmly takes root in everyday life and our consciousness. (2014, 30). Therefore search algorithms must have agency beyond their substance too, and have similarly taken root, and the interesting thing is that the algorithms are powered by people who are generating new capital, but in a radically different yet potentially just as alienating way to the industrial revolution, i.e. in terms of advertising revenue.

What is clear is that algorithms, whether as individual agents, or a part of a network, are now both ubiquitous and vital to everyday life. They effect the way

we think, work, and speak, and as well as producing and securing knowledge, they are so embedded in society that they have the ability to secure our bodies and our actions too, and that is an extraordinary power to have over us.

4.1 Language

> Neither power nor knowledge nor any other reality is anything but a mere linguistic construct." (Joseph 2006, 116)

Ever since mobile phones changed the way we spell (a semantic economy of a totally different kind), and emails knelled the death of the art of letter writing, 'technology' has been blamed for changing - perhaps spoiling - our language of communication. In this section I want to address how search engines and their algorithms are affecting language - both words already written and those yet to be imagined. What does it mean, for example, to 'natural' language, with all its richness, nuance, allusion and idiosyncrasies to be processed through binary code?

4.1.1 Programming Language

I briefly mentioned Mark-Up language in the first section when examining how categorisation and tagging can potentially limit meaning, and the important point to take from that is that in HTML, for example, the algorithm is analysing the word purely as an object which matches other pre-coded objects or, in terms of search, an object which is linked to other object-words/phrases purely by previous proximal association. As Mark Fuller and Andrew Goffey recently put it in their Evil Media manifesto:

> Harvesting data from websites is a matter of using and then stripping off the mark-up language by which web pages are rendered in order to retrieve the data of interest and returning this to a database, ready for mining. At this point, semantics is largely irrelevant. (Fuller & Goffey 2009)

Irrelevant also, as Gitelman suggests, is truth, moral judgement and accuracy (Gitelman 2013) or to use structuralist language, there is no correlation between the sign and the signified so the processing, communicating or translation of natural language by algorithm "precludes comparison of abstract notions or other objects

that do not lend themselves to direct analysis, like emotions, colors, Socrates, Plato, Mike Bonanno, and Albert Einstein." (Cilibrasi 2007) So what does this mean for our language?

> the senselessness of an object without a subject.(Fuller and Goffey Towards)

Lisa Gitelman recently raised this very point at the 2013 Governing Algorithms conference in New York. It is, she said, "a line of thinking that perhaps only an English professor could care about", but I disagree, I think it's a really important point. What happens to language when computers get hold of it? Thinking about Deleuze's writing on the inevitable hierarchies of control and access on the submission of phenomena to computation, Gitelman wonders is "the production/reduction of linguistic phenomena as/into character strings... having unnoticed intellectual effects". She continues:

> Banging ideas into character strings is simply the latest and perhaps final (hollow?) victory for a philological perspective, since online 'hits' words and phrases and sources found algorithmically by search, by term - extraction analyses, or other forms of text miningare exclusively instances with particular character strings detectable within them or indexing them, not instances with a particular logic, say, or exhibiting particular rhetorical figures, emotive appeals, or styles of argument. Rhetoric resists computation, while philology is endemic to it, if reduced to mere detectability. (Gitelman 2013)

As I argued in the first section, our online searches are only ever as good as the data available and ready for search, and the same thing goes when we reduce meaning to characters. Words, phrases, characters, novels, blogs, forums, newspapers - every available text is reduced to meaningless code. The reduction of text to code also removes the difference between media, everything digital is reduced to a binary soup, a "universal solvent into which all difference of media dissolves into a pulsing stream of bits and bytes (Lunenfeld 1999, 7).

4.2 The Commodification of Language

4.2.1 Plagiarism

Introna 2013 on plagiarism, which requires the commodification of language to exist. In ancient times poetry etc. was part of the gift economy - spoken freely and often appropriated or recited by others. The moment it becomes written in books (becomes reproducible and mobile - Introna) it enters the market and becomes a commodity. "Once a monetary value has been established, plagiarism can enter the literary discourse". (Mira See 2009, 590 in Introna 2013). Introna writes on plagiarism detection algorithms. Now language is so widely available online through search engines this adds to the problem. But importantly, plagiarism algorithms do not 'identify someone else's 'work', they merely match character strings, but in relying on the software and "incorporating these algorithms in our teaching practise we inherit this understanding and enactment of what plagiarism is - i.e. we inherit this computational reductionism - and much more besides." (Introna 2013) Introna doesn't like this as it also means we accept education as an economic exchange, a market. "If plagiarism (detected copying) is property theft then original work (not copied work or undetected copying) is property creation. Thus, this sociomaterial assemblage enacts the students as producers of commodities." (Introna 2013)

4.2.2 Adwords

I briefly touched on the idea of 'semantic capitalism' earlier, but nowhere is it more obvious than in the way Google advertising works. If you want your website advert to appear in a prominent position on the search results page, you have to bid for the words you think will best generate that search. This creates a strange world in which words have value in a way curiously abstracted and sometimes seemingly disproportionate from their usual 'ranking' of importance or relevance in everyday language and life. Google has an analytics site which suggests how much each keyword is likely to go for, and the more the advertiser wants that keyword, then the higher they bid to secure the top spot. Revenue is generated for Google on a 'pay per click' basis, the advertiser pays the bid price each time a search user clicks on their advert. The system is rigorously policed by Google, as the artist Christophe Bruno found when he tried to experiment with an Adwords 'Happening'. Bruno wanted to show how the sense and meaning of words are secondary to the commercially driven

Figure 4.1: Example of Adwords Analytics

machinations of the Web, so took out a series of small adverts which linked to his site, but instead of being keyworded adverts, he used abstract poems to see how many people actually cared about the keyword they had searched for, or whether they would let curiosity or a whim lead them somewhere else. Unfortunately for Bruno, he was unable to continue with his 'happening', as he fell foul of the strict policing of Adwords. Google don't earn from badly performing adverts, therefore if you don't try to make it as profitable as possible (by using the keywords in Bruno's case), then the results will be poor and Google will pull your advert, which is what happened to Bruno. This practise also prevents the Adword system from being used to promote things other than value goods, for example ideas, ideologies or political agendas.

4.3 Incurable Semantics

> It is wrong to reduce cultural artefacts such as literature to data (Marche 2012)

Critics like Marche (see above), have resisted the arrival of the 'digital humanities', a discipline (in both senses of the word) which Fuller and Goffey might say threatens the "incurably semantic quality of natural language" (Fuller and Goffey 2009). The reliance on 'big data' forces us to

> view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more

33

than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising it just assumed that better data, with better analytical tools, would win the day. And Google was right." (Anderson 2008)

So having created a vast database of unconnected context - a contextual soup - (i.e. the searchable database), the search algorithm now strips another layer from any kind of meaning by extracting mathematically matching data - a context less order of letter and words (Kitchin 2014) from this contextual soup. The impact of reducing literature to code is most eloquently put by Stephen Marche in his take-down of the Digital Humanities essay *Literature is Not Data*:

> Through these vast accumulations of ciphers the robots now endlessly scour for significance much the way cockroaches scour for nutrition in the enormous bat dung piles hiding in Bornean caves.... The algorithmic analysis of novels and of newspaper articles is necessarily at the limit of reductivism. The process of turning literature into data removes distinction itself. It removes taste. It removes all the refinement from criticism. It removes the history of the reception of works. (Marche 2012)

Marche makes a very important point, which I will expand on with examples in the next section: what does the algorithmic reduction of language mean to the physical canon?

Another interesting point to note is how search engines (whether keyword or semantic) will take out common words "with little semantic value" (Battelle 24), for example 'of', 'and', to', but the problem with this is although these words may have little semantic value on their own, in a sentence they are crucial to meaning.

4.3.1 Synonyms

Keyword searches using Boolean logic typically suffer from two main problems (Langville and Meyer 2012): Synonymy (where lots of words have the same meaning, for example sofa, couch, settee), and Polysemy (when one word has lots of meanings, for example bank, file, nut), which lead to potentially inaccurate search results. To combat this, most search engines now employ some kind of semantic algorithm which is meant to help interpret the context of the search and the 'intention' of the searcher and therefore define the 'relevance' of the results. Semantic

searches are usually based on Vector Space Models which turn text into numbers and then analyse the numbers to try to find context and meaning based on the already existing database of text and numbers (i.e. what is already 'on' the web - the big data). By this method (known as Latent Semantic Indexing, or LSI), the search results can be ranked on how closely the numerical versions of keywords and phrases match the precoded big databases of the web, the plan being to eliminate the problems of detextualised synonyms and polysemes.

The problem with this is that synonyms, just like the way Google translate works, are not based on meaning as such, but are calculated as 'most likely' matches from the data available on the Web. As I have already shown, if the data in the database is inaccurate, skewed or biased in any way, or privileges or discriminates against particular groups, then the semantic search result will necessarily reflect this, and not only will it reflect it, the search and any subsequent endorsement of relevance or satisfaction (i.e. a click), will reproduce, perpetuate and further embed those potentially stereotypical matches and links into the database for the next search. And so on ad infinitum. Language, and therefore meaning, is being diluted and even mutilated by the recursively of the algorithm.

There are various examples of this which have been picked up on by scholars (see Battelle 2005), including Baker and Potts, who "noticed that the term Asian was sometimes used in news articles as synonymous with Muslim" (Baker and Potts 2013, 191). The title of their paper refers to Google autopredict producing seemingly racially stereotyped results such as 'Why do White People have Thin Lips?' A recent campaign by UN Women used pictures of women with gendered and degrading stereotypical autocomplete lists across their mouths (see Fig.4.2),

but while this campaign caused outrage from popular and feminist sources who condemned Google as 'sexist', it is vital to note not only that similar searches for 'men should', 'men need' etc. produce similarly stereotypical and offensive results, but - crucially - that the autopredict function can only reflect what is on the searchable database, which, as I have shown, is not necessarily representative of society and may, due to the semantic and synonymic method of the dominant search engine, be providing an ever-expanding pool of skewed and dubious associations from which to draw.

It was exactly this problem which led me to this project topic in the first place. As part of some unrelated work, I was researching the phrase *wives and girlfriends* and how it may be considered a derogatory sexist term. I typed *wives and girlfriends*
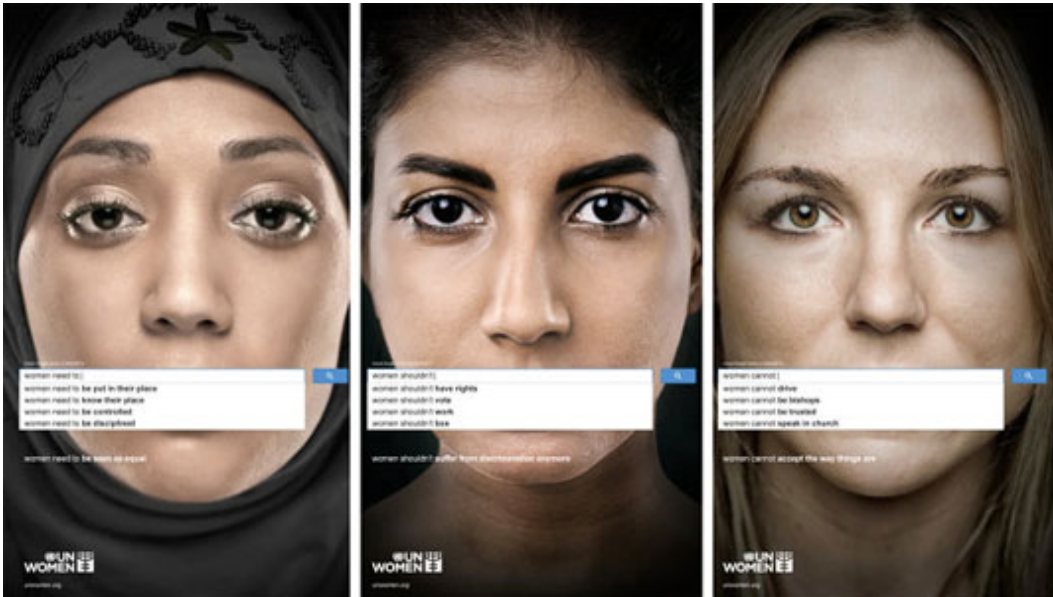
Figure 4.2: UN Women campaign against inequality 2013

sexist into Google, and was puzzled when the results came back with the word 'sexist' apparently changed to 'sexiest'. (see Fig. **??**). It was only through the research I have conducted on this project that I realised that this is not a spelling correction, or even a direct synonymic error. It is a direct result of the reliance of search engines to find semantic meaning in the commercialised, sexualised pool of decontextualized data in the searchable database.

4.3.2 Lorem Ipsum

A further example of the anomalies and linguistic and epistemological instabilities which decontextualized data can produce is an incident which took place recently regarding the place-marking cod-Latin 'language' Lore Ipsum.[1] In August 2014 researchers noticed that Google Translate was translating Lorem Ipsum text from 'Latin' to English, coming up with surprisingly modern and random results including China, NATO and Internet, which, when put together look like the hidden in plain view secret spy text from a Le Carre novel. They wondered if they had stumbled on a secret communications channel (exacerbated when the erroneous translation suddenly disappeared on the night of 16th August), or if someone had found away

---

[1]http://krebsonsecurity.com/2014/08/lorem-ipsum-of-good-evil-google-china/

Figure 4.3: Screen shot of the results of a Google search for *wives and girlfriends sexist*

to manipulate the algorithms (especially as some Lorem Ipsum translations had recently been incorporated into a DEFCON hackers challenge) , but in reality it can be explained by the way Google Translate works, which is not by actual manual translation, but by comparing already translated texts to work out algorithmically the most likely translation by word proximity. The actual meaning of words is completely irrelevant to Google Translate, it is all based on Big Data, and consequently, the more 'real' translations there are in the web archives, the more accurate the translation tends to be, so with Latin (of which Lorem Ipsum is a bastardised form), there is not much digitised data to go on (as there are not many English-Latin texts online and it is not a living, growing language), so where Lorem Ipsum is used as place-holding text on yet-to-be translated international versions of English language sites, the 'Latin' words will become confused with the original English words where there is in fact no correlation whatsoever. The only thing is does show perhaps, is a reflection of the content of the types of sites which use Lorem Ipsum as a design tool, which include, as seen in Fig. 4.4, state-owned sites such as NHS ones. While humorous alternatives to Lorem Ipsum offer such place-holding text as 'Hipster Ipsum', for example:

> Banksy gentrify pickled, deep v food truck gastropub four loko. Umami sartorial pop-up, distillery viral shabby chic brunch cray leggings vegan Bushwick roof party,[2]

this example illustrates how relying on algorithms and big data to generate and dictate meaning is actually quite dangerous and inherently insecure.

4.3.3   OCR and Print on Demand

So amongst all these glitches and problems, what good is Google Search doing? An example might be how, when researching the popular phrase 'judge a man by his questions, not by his answers', which is frequently attributed to Voltaire, and seemed extremely pertinent to this project, I was able to search for the quote in its original French which reveals its true provenance is not Voltaire but Pierre Marc Gaston Lvis in his 1810 *Maximes et rflections sur diffrents sujets de morale et de politique*, which - courtesy of the Google Book Project and the power of search, I could immediately read a scanned copy of the original. (see Fig. 4.5)
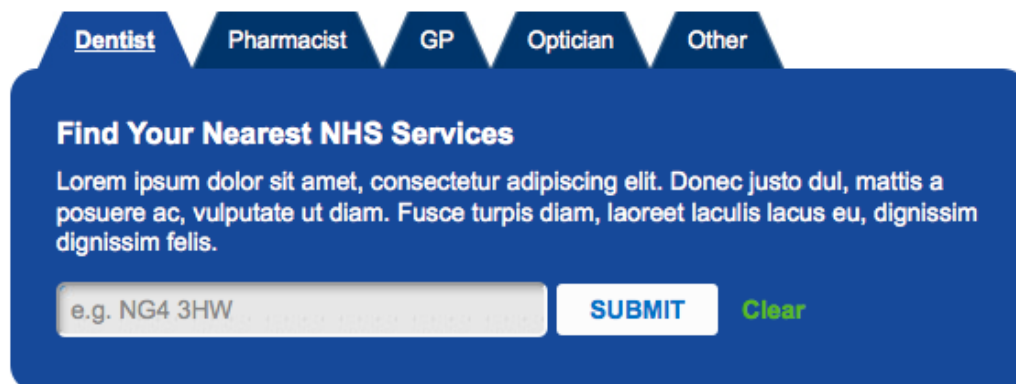
---

[2]http://hipsum.co/

Figure 4.4: Screen shot of Lorem Ipsum residue on NHS site

But other considerations of Google's ambition to digitalise the world's book, are of a bibliographic nature. Google's algorithms (with help from us, it seems), have immortalised one book over another and made that one book part of the searchable database of knowledge. But should one particular Edition now be the definitive one? What do we lose by editions (a lot in the case of the Shakespeare Folios), introductions, pictures, smells, inserts, annotations, history etc.? (see Darnton 2009). Is it making the physical library obsolete? . Does it change habits? Or maybe it makes knowledge more accessible. What about copyright? Which book to scan? is it random or will it be geographically weighted? (see Chapt.2) But for all the time Google is claiming to be doing good and making books accessible, we have to remember the links to Amazon and other book sellers on the left of the screen, or the Kindle and Google books versions. It's all revenue generating. Each time you click, you (or rather Amazon) are paying Google.

And what when the algorithms don't work so well? A personal example of this is when I bought a copy of August Strindberg's The Inferno from Amazon in 2010 during my English MA and found that the text was unreadily garbled because it was a Print on Demand version of an out of print book which had been compiled using Optical Character Recognition (OCR) software. OCR is also used by Google's reCAPTCHA program, and extension of the CAPTCHA system (Completely Automated Public Turing Test To Tell Computers and Humans Apart), which was designed to prevent sabots gaining entry to websites by asking the human user to decipher a distorted text word which the bot would be unable to do. The other side to reCAPTCHA is that the distorted words, often in an old-fashioned font, are

39

8 **MAXIMES**

XVII.

Il est encore plus facile de juger de l'esprit d'un homme par ses questions que par ses réponses.

Figure 4.5: Extract from Maximes et rflections by Pierre Marc Gaston Lvis (duc de.) 1810

words from texts which Google's OCR algorithms have been unable to read due to blurring, typeset, or annotation of the old texts which it is digitising. Any website can use the reCAPTCHA system, and the same time is helping Google to digitize the literary archive. Unfortunately in the case of my Print on Demand Strindberg text, the algorithms had not had any human assistance. (see Fig. 4.6)

What I have been trying to show with these illustrations is that algorithms need human help to accurately preserve the integrity of text. It is significant, and very obvious, says Darnton that Google does not employ a bibliographer who may be able to maintain some kind of control over the integrity of literature in a digital age and the stability of language (Darnton in Deep Search 43)

4.3.4   Autosuggest

As we have seen in the first chapter, what comes out of a search engine can only be based on what has gone into it either via previous searches or uploaded data, and this also has a direct impact on Google's Autocomplete function. According to Google "all of the predicted queries that are shown in the drop-down list have been typed previously by Google users or appear on the web". This has significant consequences. If the autocomplete is based on previous searches AND what is

Figure 4.6: Extract from Print on Demand OCR generated version of August Strindberg's The Inferno, G.P. Putnam's Sons, New York and London, The Knickerbocker Press, 1913

already on the internet then it can only reflect the content of what I have shown is a highly unstable and hierarchicised digital database. Baker and Potts argue that:

> Just over a decade into the Web 2.0 produsage model, it seems as though humans may have already shaped the Internet in their image, having taught stereotypes to search engines and even trained them to hastily present these as results of top relevance. (Baker and Potts, 2013, 201)

But this argument ignores the fact that Autocomplete is based not only on historical searches, but also on the content of the internet, as such is also necessarily shaped by the the same popular link structure as normal PageRank results, which is why an Autocomplete search returns the somewhat surprising result in Fig. 4.7. This result is not based on thousands of people typing in 'David Cameron is a willy' (as backed up the term's non-appearance as a searched keyword in Adword analytics), but is because a photograph of a banner outside a house during the election went viral on the link-heavy platform Twitter.

4.3.5   Language and Power

Fairclough's original intention when writing Language and Power was twofold: "The first is theoretical: to help correct a widespread underestimation of the significance of language in the production, maintenance, and change of social relations of power. The second is more practical: to help increase consciousness of how language contributes to the domination of some people by others, because consciousness is the first step to emancipation." (Fairclough 2001, 1) This is how I feel about the control

Figure 4.7: Screenshot taken 18/06/2014

of language which the Google algorithms have today. There are studies into how aware or conscious people are about the algorithms (Hamilton, Sandvig, Karahalios, Eslami 2014) which talk about the "curation" of newsfeeds etc. and the problems we have in becoming conscious of these power structures when "neither the user or the researcher has access to the actual proprietary algorithms at work" (Hamilton et al 2014, 631) due to blackboxing (whether this is for the protection of intellectual property or as a deliberate removal of state processes from public view "in the interests of guarding authority" (634) and potentially because the Google engineers themselves probably don't know how they work (Morozov). Their power is therefore not only beyond the consciousness and reach of the user and the researcher, but also the creator. In Fairclough's language, this is a politically very worrying barrier towards emancipation.

4.4 Security

Just like the ambiguity of the conference title 'Governing Algorithms', the term Security when applied to Search is also ambiguous. 'Feeling secure' and 'being secured' are two very different things. The recent scandal of the Facebook emotion experiment[3] has shown how much power algorithms and personal data can have over people in a physical, very real way - not just potentially - and, although Facebook claimed permission for such experiments is a term of use, there is no way of knowing whose data and emotions were manipulated and the consequences of such manipulation. How many of those participants would have been willing if they had actually been aware of the experiment? There are many different concepts and per-

---

[3]When an academic paper was published detailing an experiment where Facebook newsfeeds were covertly doctored to see if it affected the emotional content of future posts

ception of security, from the theoretical to the technical and from a wide political spectrum. It needs more space than this to expand, but the crux of the question is - how much (potentially comprisable) data do we want to give away in return for the benefits of search, which include ontological security, satisfaction of having all the knowledge, confidence of knowledge, security in knowledge.... and the other positive things which come out of search such as medical diagnostic and drug research (Tufekci 2014, 16). The Facebook incident has shown how personal data can be used to manipulate emotion, which leads to the question: could this technique be used for criminal or subversive purposes? Even if Google does not conduct experiments like the Google one, if our data is not secure (from either state or non-state actors), there is the chance that it could be used to sway opinion or as propaganda. There is no getting away from our intentional or unintentional interactions with Google, but have we been empowered and liberated by the part we now play in the production of data (Web 2.0 ref?) and the democratised web (Beer 2009) or are we controlled by it? Where is the power in social participation, asks Beer (2009). What price do we pay for it? The perfect search engine requires a "Faustian bargain" of exchange of personal data for knowledge, says Zimmer (Zimmer in Spink and Zimmer 2008, 93). Although the line between state and non-state surveillance through media such as Google has been blurred in light of the Snowden revelations, the big difference is motive, which in the case of Search is, as we have seen, purely financial rather than philanthropic, utilitarian or for the good of society. Rohle distinguishes between forms of power in search engines, seeing a disciplinary punishment/reward kind of power between Google and the webmasters, and a power through a less invasive knowledge and collection of personal data between Google and the search users. (Rohle in Deep Search 118)

4.5   Surveillance by Algorithm

For the purposes of this section I will use David Lyon's popular definition of surveillance which lends itself well to the study of algorithms and online search (Zimmer 2008). Surveillance, according to Lyon, is any collection and processing of personal data, whether identifiable or not, for the purposes of influencing or managing those whose data have been garnered (Lyon 2001, 2). Synthesised into a process of 'dataveillance' (Clarke 2008), this section will unpick the salient keywords in Lyons' definition.

4.5.1 Data Collection

When Google started scanning Gmail emails for advertising purposes there was outrage, but Google insisted it was all okay because human beings didn't read the emails, just machines, so worrying about a computer reading your email is like worrying about your dog seeing you naked. Much the same argument surfaced after the Snowden revelations regarding the NSA and GCHQ mass harvesting data from Google, Yahoo etc., (and there were strongly denied accusations of these companies providing 'backdoor' access to state agencies) - where the issues lay around the definition of collection. Data was not deemed to have been collected until it had been "processed into intelligible form" and "received for use by an employee of a DoD", so they were able to say (as Clapper did to the Senate) that because the mass of the data isn't seen by a human unless specifically accessed under legal process, it hadn't officially been collected. I think in these two circumstances Google's explanation actually holds more water than the NSA's, because of the obfuscated means of data collection, the blackboxed technology and the secrecy of the agency accessing the data, it is impossible for a data subject to know exactly who sees that data, how secure it is from criminals, hackers, advertisers, or other law enforcement agencies. It can, writes Battelle, be "discovered subpoenaed, archived, tracked, and exploited for all sorts of ends" (Battelle ?), whereas, as Bruce Schneier so eloquently puts it there isnt a court order in the world that can get that information out of your dog.

4.5.2 Identifiable Personal Data

It is common these days to read news stories which detail incriminating online searches carried out by convicted criminals.

> Etched into the silicon of Google's more than 150,000 servers, more likely than not, are the agonized clickstreams of a gay man with AIDS, the silent intentions of a would be bomb maker, the digital bread crumbs of a serial killer. Through companies like Google and the results they serve, an individual's digital identity is immortalized and can be retrieved upon demand. (Battelle 2005)

As hard as Google say they protect our data, the very fact that they have it in the first place and it is potentially accessible, is enough to meet the definition of

surveillance, whether we are actually being 'identified' by name or not.[4] If logged in to a Google service, all search data will be linked to actual personal details (or those of the holder account anyway), which may include links to social media, telephone numbers and even addresses. All this is linked together to create a self generated persona which may or may not be the same as a person's 'real identity' or "new algorithmic identity" (Cheney-Lippold 2011). The point made by Lyons' phrase is that it doesn't actually matter whether a person can be identified or not by their data - the fact that the data is being collected alone makes it surveillance. In fact, to take that further, the fact that the data *could* be collected is enough to make it surveillance, an algorithmic gaze which has

> no need for arms, physical violence, material constraints. Just a gaze. An inspecting gaze, a gaze which each individual under its weight will end by internalizing to the point that he is his own overseer, each individual thus exercising this surveillance over, and against, himself. A superb formula: power exercised continuously and for what turns out to be a minimal cost." (Foucault 1980, 155)

It is not hard to see the potential leverage power which possession (or potential possession) of such personal data could have in terms of blackmailing, bribery, corruption, or even in terms of security vetting - it may seem like a dystopian nightmare, but the already intrusively personal demands made by government departments to 'clear' staff to high levels of security clearance could be made a whole lot more personal with search data, which crucially wouldn't actually need to be collected - the threat that it *could* be used (maybe with the candidate's permission), would be enough to control the behaviour of a candidate either in their future search activities or in their decision to apply for vetting in the first place. Of course this type of surveillance is evidentially problematic in terms of who uses what devices, security of personal accounts and even in terms of search queries which, with au predict, can sometimes send the user in a direction they may not have been intending and did not actually search for, (or intended to but changed their mind as in the case of key-logging), but in terms of 'intelligence', especially building up a picture of someone, search data is an incredibly powerful tool. But even not using Google, for example, wouldn't keep you 'safe' from surveillance by algorithm:

---

[4]see NY Times article on Google passing details of algorithmically detected child pornography traffic to police: http://bits.blogs.nytimes.com/2014/08/04/google-gives-child-pornography-email-evidence-to-police/

Even a total refusal to use Google would offer no escape. So many
others use it (for example in the form of its Gmail email service) that a
substantial amount of one's personal email communications will end up
in its domain. And so many websites and services are interwoven with
its various offerings that it is unavoidable." (Becker and Stalder 2009, 7)

4.5.3 Biopower

With Foucault deployed it is but a short step to Biopower. From the Lyons quote, it
it suggested that surveillance is for the purpose of "influencing and managing those
whose data have been garnered." If we take a Foucauldian reading of the panopti-
conic power of the search algorithm, we can see how its implicit omniscience and
omnivision might produce exactly such a self-policing, self regulating, internalised
self-security, a potential power of concern in the hands of the state and its intelli-
gence agencies, but even more so in control of a private, unregulated, commercial
company such as Google. Exposure to the potential threat of permanent visibility
will "induce in the inmate a state of conscious and permanent visibility that assures
the automatic functioning of power (Foucault 1977, 197). So we have a "a rapidly
emerging soft cage of everyday digital surveillance" (Zimmer in Spink and Zimmer
2008, 77), which is all too easily transferred to physical 'cages' and spaces in terms
of how code and software can 'sort' us into geographic spaces too (see Graham on
Software Sorting and Kitchin and Dodge Code/Space).

4.6 Filter Bubbles and Echo Chambers : Self Security

In 2011 Eli Pariser published a hugely successful book called *The Filter Bubble:
What the Internet is Hiding from You*, which advanced the theory that online per-
sonalisation is limiting, not expanding, our access to information as Google seeks
to return searches based on previous interests and likes, measured by what it per-
ceives as previous 'successful' results. The problem Pariser identifies is that we are
in danger of never seeing any information which challenges our opinions, and us or
is outside of our comfort zones or social circles. Also called the "echo chamber"
effect, this phenomena, when amplified to a political level, has serious implications
for the operation of a successful democracy, which relies on the concepts of debate
and access to all side of information to succeed. Diaz cites JS Mill on this subject:

He who knows only his own side of the case knows little of that. His reasons may be good ... but if he is unable to refute the reasons of the opposite side, if he does not so much as know what they are, he has no ground for preferring either opinion ... [H]e should hear the arguments ... from the persons who actually believe them, who defend them in earnest and do their very utmost for them. He must know them in their most plausible and persuasive form. (35)

4.6.1 Democracy

Personalisation has created online personas which feed purely from themselves and from those who agree with them, even if the algorithms can sometimes get this completely wrong, (see Jarrett on "intentional fallacy"). Further to this, a false reliance that the algorithm 'knows what's best for you', can lead to over reliance on their choices, whether it is what film to watch, or, more seriously, who to vote for. As Ford writes:

Why should I vote for political candidates myself when my cyber doppelganger, armed with the voting records, every public statement, policy paper and publication of every candidate and my preference profile, can choose the best candidate for me. (Ford 2005 1579)

And why stop there? Why not let the algorithms vote for everybody? (see discussion of Aaronson's Eigendemocracy). John Danaher refers to this frightening prospect of the potential political power of personalisation as 'algocracy', (John Danaher Rule by Algorithm), while Princeton computer science professor and cyberactivist Ed Felten was the 'Googlocracy' (2004) as a potentially successful voting system.

In 2010 Robert Epstein and Ronald E. Robertson conducted controlled experiments to see if search rankings could sway voting preferences without voter awareness (using 2010 Australian election). They concluded that "unregulated search rankings could pose a significant threat to a democratic system of government." (Epstein and Robertson 2013)

On the other end of the democratic spectrum, in a recent article entitled 'Mouth breathing Machiavellis Dream of a Silicon Reich', journalist Corey Pein wrote about a seemingly growing trend amongst the ranks of the technical elite of what he

calls the "neoreactionary vanguard of the dark enlightenment", a number of whom (including Google employee Justine Tunney) have publicly denounced democracy in favour of non-elected corporate governance led by the tech industry.

4.6.2 Ontological Security

The search for answers is inherently tied up with the need to feel ontologically secure, sure of yourself and secure within your environment.

beginquote To be ontologically secure is to possess, on the level of the unconscious and practical consciousness, answers to fundamental existential questions which all human life in some way addresses (Giddens 1991, 47) endquote

As we have seen with features such as Autocomplete, there is a certain amount of security which can be gained from thinking that other people have had the same questions or problems as you, but this can also lead to kind of confirmation bias which - like the Filter Bubble - rewards similarity and not difference. Sanz and Stancik have recently tried to fill what they see as a gap in the research on search in relation to ontological security, i.e. the cultural implications as opposed to the discourses of power and economics which they say have been written about a lot. But this will only work if they are examining the actual search engine and algorithms and the structures behind that, as well as the results and how they shape societies culturally. They say "online search offers a unique empirical window into the study of culture.". (Your search Ontological Security matched 111,000 documents: An empirical substantiation of the cultural dimension of online search 2014, 253), but I think I have shown how problematic a purely empirical study might be.

Chapter 5

To the Future

Back in 2000 Introna and Nissenbaum wrote an incredibly prescient paper on the perceived and potential power they saw algorithms having over society and the web. They called for increased regulation and algorithmic transparency, a"full and truthful disclosure", to enable a fair and democratic search engine system, arguing that the potential power then given to spammers and gamers (the oft-cited defences given for black-boxing the search algorithms) would be diluted and itself democratised and made fairer by everybody having access to the information. This would also open up a more competitive market in search engines. (Introna 2000, 61). The problem with that is that Google (and other Search Engines) remains a private company with no state regulation or social obligations beyond its market-infused rhetoric, and therefore has no reason to desire an egalitarian Web. But, as Granka suggests, "online information is too important and too special a commodity to be shaped by market forces alone", (Granka 365) and over the last few years this resistance to the algorithmic and corporate shaping of knowledge has led to academic, technical and creative (and mixtures of all) challenges, responses and interventions which have sought either to test or expose the hidden power of the algorithms using a variety of innovative and creative research methods, or to work with the algorithms in a bid to subvert their control. As Mane and Uprichard conclude, only thinking of algorithms in terms of control, surveillance and hierarchy is to submit to that control (260).

5.1 Taking Back Control

"Stop searching, start questioning", Geert Lovink urged us in his Society of the Query essay (Lovink 2008), pointing out that despite the monopolised power of the search engine, Web users have "such a multitude of tools at their disposal to distribute power." (Lovink 2008,1), and after all there are *other* search engines. Perhaps it is a matter of learning to use some of those tools better, whether auto-didactically or through greater engagement with the grammars and mechanics of

digital technology through education, as a recent call for the re-politicisation and mandatory teaching of code writing in schools shows. http://codeactsineducation.wordpress.com/2014/09/0 computing-curriculum/ As well as the raft of popular literature on search engines and Google (Levy, Pariser, Morozov), there are also scholars trying to bridge the gap with more scientific, but still accessible, semi-academic books (Langville and Meyer). As we will see in the next section, (and have already seen in light of security implications) some scholars and activists are beginning to realise that the generative nature of the Web provides the interactive user with considerable powers to influence content, especially collectively. "If we understand algorithms as mirrors, we can actively take on responsibility to create our own mirror image. We have never had that chance before. We have never been that kind of enabled citizen. Lets stop measuring Google against some sort of mythical normative standard and start to think about how we can use Google creatively." (Mahnke and Uprichard 2014, 267)

5.1.1 New Methods

Several scholars have been experimenting with new research methods to understand and mitigate the effects of algorithms (Sandvig, Feuz). New methods are urgently needed, otherwise, as Martin Feuz fears "the knowledge and power differentials between those on the inside of search engines and those who are mere users of a powerful but opaque machine are bound to grow." (Feuz 2011)

Crawford urges less reliance on big data as a stand-alone research tool: "Longer term, we must ask how we can bring together big data approaches with small data studies computational social science with traditional qualitative methods. We know that data insights can be found at multiple levels of granularity, and by combining methods such as ethnography with analytics, or conducting semi-structured interviews paired with information retrieval techniques, we can add depth to the data we collect." (Crawford 2013), and Rob Kitchin has highlighted how important it is to recognise the different objectives and motives of business and the academy in working out how to make sense of big data and algorithms (Kitchin 2014).

5.1.2 A New Knowledge Logic

What is evident is that the technologies of the Web are now inextricable from and actively shape society and daily life. Search engines in particular control access to

information and therefore contribute to the social construction of knowledge (LM Hinman 67), but this process is complicated by the commercial and private nature of the companies who play a part in such construction. As Hinman warns, "the keys to the kingdom of knowledge have been passed to a for-profit company whose system of ranking is a closely-guarded trade secret." (Hinman 70) Lawrence Hinman on the "Searching Ethics: The Role of Search Engines in the Construction and Distribution of Knowledge" (chapt 5 in Spink and Zimmer 2008).

As the technological discourse changes, so will the discourse of society (Scollon and Scollon, 2004, p. 7), and these new knowledge logics (Gillespie 2014) need to be studied. As research fields expand in time with technology, there have been calls for new academic disciplines such as Internet Studies, or even Web Search Studies (Zimmer 2010), which would facilitate a more representative and appropriate depth of study than the already bursting Digital Humanities can accommodate, and centres like the Oxford Institute for Internet Studies and the Cambridge Centre for Digital Knowledge are being created to study digital epistemologies and "how new knowledge is produced in the digital domain." http://www.crassh.cam.ac.uk/programmes/cambridge-centre-for-digital-knowledge

5.2 Alternatives

So what are the alternatives? Armed with new tools and knew knowledge, what can be done to "tame the algorithmic tiger" (Mahnke and Uprichard 2014), 257)?

5.2.1 Transparency

Google's search algorithm is a closely guarded secret for commercial reasons, but Gillespie points out that other platforms such as Reddit let the user play with the parameters such as 'hot topics' and 'trends' to see how the results change (this is from Gillespie interview 2014). In fact even Google has provided users with tools such as Google Trends, Analytics and Keywords in Adwords, which give some insight into the workings but are in effect little more than appeasement without the workings inside the black box which, as Morozov speculates, are probably far too convoluted and complicated for any Google engineer to comprehend anyway. (Morozov).

5.2.2 Competition

How about not bursting the filter bubble, but allowing competition between lots of bubbles (see Crawford on agonistic pluralism)? She advocates an "agonistic ideal, where public relevance algorithms could acknowledge difference and dissent rather than a silently calculated public of assumed consensus and unchallenged values." (crawford 2013)

5.2.3 Exploring Links

PageRank isn't the only way to rank search results, and research collective Metahaven have been using Graph Theory to explore different link-based structures of ranking which return results which may be on the periphery, but have a wide range of links from wide range of sources. (Metahaven in Deep Search 2009)

5.2.4 Regulation

Search is not a public utility, but there is growing debate about whether it needs to be regulated as such (see Jamison 2012). Introna and Nissenbaum led the call for regulation in 2000, but there are always the warnings of Orwellian state power dismissing such calls, despite just those powers currently being in the hands of a less accountable private company. Perhaps it is just a matter of the speed of the growth of the Web in general, and it just seems so unwieldy because "few laws have caught up with the technology. http://www.pcadvisor.co.uk/features/internet/3304956/how-google-facebook-and-amazon-run-the-internet/ Baker and Potts (2013) suggest there should be an option to flag 'offensive' autopredict results, but this just opens up to same arguments in terms of demographic of usership and different people, and what is offensive etc.

5.3 Interventions

The increasing ubiquity of Google as the world's knowledge provider and the ever expanding databases of available information have led not only to a wealth of multi-disciplinary research and academic exploration, but has also provoked reaction in experimental artistic, creative and perhaps more pragmatic approaches. People and groups who have tried to harness the vastness and fluidity of the information processed by Google in alternative ways and for other - no less important - reasons.

Lovink concludes that "it's not the glut of information that is the problem, but our ability to use it intelligently... we need to invent new ways to interact with information, new ways to represent it, and new ways to make sense of it. How are artists, designers, and architects responding to these challenges?" (Lovink 2008, 9)

5.3.1 Testing the Algorithm

Trying to reverse engineer blackboxed algorithms is methodologically challenging in many ways, not least due to ethical considerations (Sandvig), but because the information on the Web is inherently unstable and malleable, and because personalised search has made it near impossible to collect empirically sound, geographically and temporally stable data (Feuz).

Despite this, several scholars have had a go, including Martin Feuz, Matthew Fuller and Felix Stalder who set up a test study Feuz et al. (2011) by creating fictional Google accounts for the philosophers Kant, Nietzsche and Foucault and creating their online 'personas' by feeding the indexes of their works into the Google Search Engine and conducting controlled searches from each account. Unsurprisingly, the study showed how the search results for the three accounts fluctuated exponentially as the personalisation took hold and that the benefits of personalisation serve the advertisers more than the users. In a case of art imitating life, Feuz has since imagined a day when such 'search history personas' might be traded as commodities on EBay.

5.3.2 The Reclamation of Serendipity

Much criticism of personalised search results focuses on the negative effects of losing the element of random luck which might send you in an unexpected but fortuitous direction. This has been referred to as "the death of serendipity", and is reflected in the hierarchical tracts of knowledge which undermine the randomness of the rhizomatic structure of the Web(Hess in Spink and Zimmer 2008, 41). But there ways to counter this effect such as new data mining software like Ayasdi which claims to provide what Clark terms 'digital serendipity'. (Clark 2013) http://www.wired.co.uk/news/archive/2013-01/16/ayasdi-big-data-launch

It might also be useful to think about the 'new serendipities' which might arise from the quirks of search, or the "certain sense of the tendentially absurd" which can appear in software (Feuz 2011), something which Fuller and Goffey explore as

one of their Strategems of Evil Media Studies - "stratagem five: make the accidental the essential" (Fuller and Goffey Towards an Evil Media Studies)

5.3.3   Ludic Subversion

With its title borrowed from Lovink's 2008 essay, the Institute of Network Cultures series has recently published a Reader called Society of the Query, which brings together a range of recent research with an aim to making the invisible workings and influences of search algorithms visible, just as Debord and the Situationists tried to expose the Spectacle. Their main aim is "to design visibility campaigns to make their influence apparent". (Konig and Rasche 2014, 10) The title is deliberately political and the book plays on Derrida's Archaeology of the Frivolous (1973) to look at ways looking "to disrupt and disturb and deconstruct that which is taken for granted."

The editors Mahnke and Upritchard imagine what would happen if you could get billions of users to enter specific search terms which 'correct' misconceptions or stereotypes in the database which appear through Autopredict; whether there is a positive power for change in the algorithms. "Perhaps the capacity to act collectively towards a common good is by actively interacting with Googles autocomplete in ways that are similar to activist movements. If Google can shape our views, perhaps we can use Google to change our views?" (Mahnke and Uprichard 2014, 266), although unfortunately it seems unavoidable that any attempt to 'change normative attitudes' in this way will itself be too situated and nuanced a method to avoid criticism.

However, in a recent essay on the power of Facebook, Peter Olsthoorn also seeks to harness the power of playful subversion and disobedience: "Through data collection, Facebook and Google are the most omniscient market research bureaus in the world at present. Of course this will only remain so if we remain faithful to it, searching and commenting in the relevant boxes, like good boys and girls." Peter Olsthoorn http://www.netkwesties.nl/716/the-facebook-machine-and-the-power.htm 17/08/2014.

5.4   Artistic Interventions

Of course interventions don't have to be overtly manipulative and participatory to have effect, and there are many examples of "software art" (Feuz 2011) which show this.
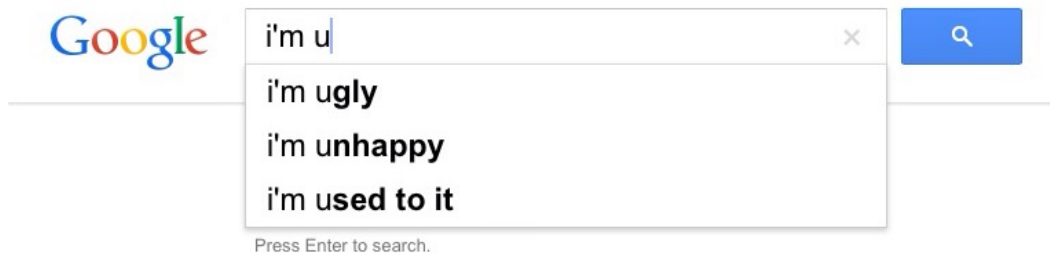
Figure 5.1: example of Google Poetics 03/08/2014

5.4.1   Google Poetics

Since 2012 Social Media Producer Sampsa Nuotio has been running a website called Google Poetics (www.googlepoetics.com) with it's excellent tag line "Google writes poetry on subjects that people are truly interested in." As Martina Mahnke and Emma Uprichard point out: "algorithms can instigate and facilitate imagination, creativity, and frivolity, while saying something that is simultaneously old and new, always almost repeating what was before but never quite returning." (Mahnke and Uprichard 2014), and while there is something of the absurd about the poems, they is also something of the poignant. For whatever skewed or stereotypical reason, the strings of words which pop up *are* phrases which exist online. Somebody has put them there so a certain extent they create a sad reflection of life. (Fig. 5.1)

5.4.2   The Human Browser

The French artist Christophe Bruno has conducted similarly linguistic interventions which have aimed at "diverting global symbolic structures like Google search engine" (Bruno Interview 2006). The Human Browser is a kind of mobile installation during which an actor is fed and articulates a text-to-voice stream of search results which are generated by the environment, situation or space through which the actor moves. Bruno's point is to show how speech and sensical language has become subsumed by the ubiquity and invasiveness of the Web, and as lost its individual power and agency. In fact only now in the hijacking and subversion of the global structures of power which make up the Web, is language effectively 'effective'. (Bruno interview 2006)

5.5 Difficulties of Intervention

As already discussed, search personalisation has made it methodologically difficult to study search engines, as there is effectively no control or baseline to work against (Feuz et al. 2011) and added to the problem of curated and ever changing results and ever-tweaked algorithms, is the fact that at any given time Google or other search engines could be conducting A/B type controlled test experiments on users. The recent furore over the Facebook 'emotion' experiments show the ease and propensity with which users can be experimented on, and Google are constantly assessing and tweaking their algorithms to maximise profit through 'successful' clickstreams.

But as well as monitoring for improvements, the Google crawlers are constantly on the look out to block spam and automated searches and 'link farms' which are used to boost search rankings, which means that legitimate research or investigation gets caught up in these purges too. Martin Fezs philosopher experiment , for example ran into trouble when Google blocked several IP addresses they were using assuming that they were processing automated requests from a virus or spyware application. Similarly, Christophe Bruno's Adwords 'Happening', during which he experimented with buying keywords which linked to abstract poetry was taken down by Google because it perceived that the adverts were not being successful enough as Google polices Adwords quite stringently to make sure that commercial profit is the *only* motive for the purchase of keywords, rather than ideological or political motivations (Lessig 2004, Diaz 2008). Another example of algorithms thwarting research came from Facebook recently, when someone tried to see what would happen if they 'liked' everything which appeared on their newsfeed. Apart from the worrying speed with which all human content disappeared from his timeline in favour of more of the adverts which he was 'liking', and the extreme ideological swings he saw in his newsfeed after 'liking' particular stories, his experiment was fairly quickly outed by the Facebook algorithms as non-normative behaviour and he was contacted by their PR department. It does make you wonder if it is actually possible to research such algorithms from outside of their normative bubbles.

5.6 Towards a Theory of Reconstruction

As I have shown, algorithms have considerable agential power both in terms of security and society. My own intervention has been to start thinking about a theory

of Reconstruction.

Since Derrida and the post-structuralists revealed the embedded codes which linger in language and imbue society with rigid binary systems of structure, the social sciences and the humanities have been eager to deconstruct these restrictive and unimaginative linguistic rules.

In Big Data and the algorithmic reproduction of information, knowledge and language, I would suggest that the social and political discourses are being re-coded with exactly the type of reductive structures which we have tried for decades to expose and tame. I therefore propose a new critical theory of Reconstruction, which I hope will form the basis of my PhD thesis.

.

Chapter 6

Conclusion

This project has set out to examine online search and its algorithms from three interlinked and interdependent angles, the database, the workings of the algorithm (as much as we can see them), and the effects that search has on securities, societies and language, all of which are, I believe dangerously reliant on the benevolence of the Search Engine companies and in particular Google, for the ongoing security of both our data and knowledge.

The side-effects of search are far reaching, and potentially devastating, but can perhaps be mitigated by regulation, greater transparency, or even subversion, their power being apparently "mutable, once we start playing around."(Konig 259).

Politically, we have seen that the seemingly liberal, and democratising Web is complicated and problematised by the processes and effects of search on freedoms and securities of self and of language.

Kitchin tells us that "Coding is a disciplinary regime", (Kitchin and Dodge Code Space), and it does indeed seem to be the case that in the post-hegemonic (Beer/Lash) age of technology there seems to have been a vacuum of power which has been filled by a commercially driven, unaccountable private company which can perhaps only be checked by a reversal of the panoptic gaze.

Bibliography

Aaronson, S. (2014). Eigendemocracy. http://www.scottaaronson.com/blog/?p=1820. Shtetl-Optimized.

Amoore, L. (2006). Biometric borders: Governing mobilities in the war on terror. *Political geography*, 25(3):336–351.

Amoore, L. (2014). Security and the incalculable. *Security Dialogue*, page 0967010614539719.

Anderson, C. (2008). The end of theory. *Wired magazine*, 16(7):16–07.

Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Penguin.

Becker, K. (2009). The power of classification: Culture, context, command, control, communications, computing. *Deep search: The politics of search engines beyond Google*, pages 163–172.

Borges, J. L. (1964). *Labyrinths: selected stories & other writings*, volume 186. New Directions Publishing.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Bruno, C. (2002). The google adwords happening. http://www.iterature.com/adwords/. http://www.christophebruno.com/.

Carr, N. (2008). Is google making us stupid? *Yearbook of the National Society for the Study of Education*, 107(2):89–94.

Cheney-Lippold, J. (2011). A new algorithmic identity soft biopolitics and the modulation of control. *Theory, Culture & Society*, 28(6):164–181.

Crawford, K. (2013). The hidden biases in big data. *HBR Blog Network*, 1.

Dalton, C. and Thatcher, J. (2014). What does a critical data studies look like, and why do we care? seven points for a critical approach to big data. *Society and Space Open Site.*

Darnton, R. (2009). The library in the information age: 6000 years of script. *Deep search: The politics of search engines beyond Google*, pages 32–44.

Diaz, A. (2008). *Through the Google goggles: Sociopolitical bias in search engine design.* Springer.

Directive, D. (1982). 5240.1r,". *Activities of DoD Intelligence Components that Affect US Persons," November*, C2.2.1.

Epstein, R. and Robertson, R. E. (2013). Democracy at risk: Manipulating search rankings can shift voters preferences substantially without their awareness. In *25th annual meeting of the Association for Psychological Science.*

Fairclough, N. (2001). *Language and power.* Pearson Education.

Feuz, M., Fuller, M., and Stalder, F. (2011). Personal web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation. *First Monday*, 16(2).

Finkelstein, S. (2008). Google, links, and popularity versus authority. *The hyperlinked society: Questioning connections in the digital age*, pages 104–120.

Foucault, M. (2012). *The archaeology of knowledge.* Random House LLC.

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Fuchs, C. (2011). A contribution to the critique of the political economy of google. *Fast Capitalism*, 8(1).

Fuchs, C. and Horak, E. (2008). Africa and the digital divide. *Telematics and Informatics*, 25(2):99–116.

Fuller, M. and Goffey, A. (2009). Toward an evil media studies. *Parikka and Sampson*, 149.

Fuller, M. and Goffey, A. (2012). *Evil media*. MIT Press.

Galloway, A. R. (2012). *The interface effect*. Polity.

Gandy Jr, O. H. (1996). Coming to terms with the panoptic sort. *Computers, surveillance, and privacy*.

Gerhart, S. (2004). Do web search engines suppress controversy? *First Monday*, 9(1).

Gillespie, T. (2014). The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, page 167.

Goldstein, R. (2014). *Plato at the Googleplex: Why Philosophy Won't Go Away*. Random House LLC.

Graham, M. (2014). Geotagging reveals wikipedia is not quite so equal after all. http://theconversation.com. Accessed 18/08/2014.

Graham, S. D. (2005). Software-sorted geographies. *Progress in Human Geography*, 29(5):562–580.

Granka, L. A. (2010). The politics of search: A decade retrospective. *The Information Society*, 26(5):364–374.

Grimmelmann, J. (2008). Google dilemma, the. *NYL Sch. L. Rev.*, 53:939.

Hamilton, K., Karahalios, K., Sandvig, C., and Eslami, M. (2014). A path to understanding the effects of algorithm awareness. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 631–642. ACM.

Hess, A. (2008). Reconsidering the rhizome: A textual analysis of web search engines as gatekeepers of the internet. *Web search: Multidisciplinary perspectives*, page 35.

Hinman, L. (2008). Searching ethics: The role of search engines in the construction and distribution of knowledge. *Web Search: Multidisciplinary Perspectives*, 14:67.

Introna, L. (2013). Algorithms, performativity and governability. http://governingalgorithms.org. Presented at Governing Algorithms Conference New York May 2013.

Introna, L. D. and Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The information society*, 16(3):169–185.

Jarrett, K. (2014). A database of intention? *Society of the Query: Reflections on Web Search*, pages 16–29.

Jeanneney, J.-N. (2008). *Google and the myth of universal knowledge: A view from Europe.* University of Chicago Press.

Jenks, C. (2002). *Cultural reproduction.* Routledge.

Joseph, J. E. (2006). *Language and politics.* Edinburgh University Press.

Kitchin, R. and Dodge, M. (2006). Software and the mundane management of air travel. *First Monday.*

Kitchin, R. and Dodge, M. (2011). *Code/space: Software and everyday life.* MIT Press.

Klauser, F., Paasche, T., and Söderström, O. (2014). Michel foucault and the smart city: power dynamics inherent in contemporary governing through code. *Environment and Planning D: Society and Space*, 32:000–000.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Lee, M. (2010). A political economic critique of google maps and google earth. *Information, Communication & Society*, 13(6):909–928.

Lovink, G. (2008). The society of the query and the googlisation of our lives a tribute to joseph weizenbaum. *Eurozine*, 2008.

Lovink, G. (2009). Society of the query: The googlization of our lives. *Deep search: The politics of search engines beyond Google*, pages 45–53.

Lyon, D., editor (2003). *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination.* Routledge.

Mackenzie, A. (2006). *Cutting code: Software and sociality*, volume 30. Peter Lang.

Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. *Information, Communication & Society*, 15(5):769–787.

Morozov, E. (2012). *The net delusion: The dark side of Internet freedom.* PublicAffairs.

Morozov, E. (2013). The real privacy problem. *TECHNOLOGY REVIEW*, 116(6):32–43.

Neyland, D. (2014). On organizing algorithms. *Theory, Culture & Society*, page 0263276414530477.

Pasquinelli, M. (2009). Googles pagerank algorithm: A diagram of cognitive capitalism and the rentier of the common intellect. *Deep Search*, 3:152–162.

Ragnedda, M. and Muschert, G. W. (2013). *The digital divide: The internet and social inequality in international perspective*, volume 73. Routledge.

Reilly, P. (2008). Googling terrorists: Are northern irish terrorists visible on internet search engines. *Web search: Multidisciplinary perspectives*, page 10.

Rieder, B. and Sire, G. (2013). Conflicts of interest and incentives to bias: A microeconomic critique of googles tangled position on the web. *New Media & Society*, page 1461444813481195.

Rimington, S. (2001). *Open secret: The autobiography of the former director-general of MI5.* Random House.

Rohle, T. (2009). Dissecting the gatekeepers: Relational perspectives on the power of search engines. *Deep search: The politics of search engines beyond Google*, pages 117–132.

Rose, G. (2014). Interface, network and friction: cultural geography and digital technologies. http://conference.rgs.org/AC2014/224. Plenary presentation given at Royal Geographic Society annual conference 28/08/2014.

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. http://www-personal.umich.edu/ csandvig/research/. Paper presented to 'Data

and Discrimination: Converting Critical Concerns into Productive Inquiry' a pre-conference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle USA.

Sheff, D. (2004). Playboy interview: Google guys. *Playboy*, 51(9):55–60.

Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778.

Spink, A. and Zimmer, M. (2008). *Web search: Multidisciplinary perspectives*, volume 14. Springer.

Stalder, F. and Mayer, C. (2009). The second index: Search engines, personalization and surveillance. *Deep search: The politics of search engines beyond Google*, pages 98–115.

Toffler, A. (1981). The third wave, chapter the rise of the prosumer.

Tufekci, Z. (2014a). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19(7).

Tufekci, Z. (2014b). What happens to #ferguson affects ferguson: Net neutrality, algorithmic filtering and ferguson. https://medium.com/message/ferguson-is-also-a-net-neutrality-issue-6d2f3db51eb0. https://medium.com/ 14/08/2014.

Vaidhyanathan, S. (2012). *The Googlization of everything:(and why we should worry)*. Univ of California Press.

Van Couvering, E. (2010). *Search engine bias: the structuration of traffic on the World-Wide Web*. PhD thesis, London School of Economics and Political Science (LSE).

Vaughan, L. and Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information processing & management*, 40(4):693–707.

Weber, S. (2007). Das google-copy-paste-syndrom. *Wie Netzplagiate Ausbildung und Wissen gefahrden. Telepolis*.