- About
- Contact

Home » Blogs » felix's blog

The Second Index. Search Engines, Personalization and Surveillance (Deep Search)

Introduction[1]

Google's well-known mission is "to organize the world's information". It is, however, impossible to organize the world's information without an operating model of the world. Melvil(le) Dewey (1851-1931), working at the height of Western colonial power, could simply take the Victorian world view as the basis for a universal classification system, which, for example, put all "religions other than Christianity" into a single category (no. 290). Such a biased classification scheme, for all its ongoing usefulness in libraries, cannot work in the irreducibly multi-cultural world of global communication. In fact, no uniform classification scheme can work, given the impossibility of agreeing on a single cultural framework through which to define the categories.[2] This, in addition to the scaling issues, is the reason why internet directories, as pioneered by Yahoo! and the Open Directory Project (demoz)[3], broke down after a short-lived period of success.

Search engines side-step this problem by flexibly reorganizing the index in relation to each query and using the self-referential method of link analysis to construct the ranking of the query list (see Katja Mayer in this volume). This ranking is said to be objective, reflecting the actual topology of the network that emerges unplanned through collective action. Knowing this topology, search engines favor link-rich nodes over link-poor outliers. This objectivity is one of the core elements of search engines, since it both scales well and increases the users' trust in the system.

Yet, this type of objectivity has its inherent limits. The index knows a lot about the information from the point of view of the providers who create the topology of the network (through creating links), but knows nothing about the searcher's particular journey through the informational landscape. What might be relevant on an aggregate, topological level could well be irrelevant on the level of the individual search interest. This problem is compounded by the fact that the actual customers of the search engines, the advertisers, are not interested in the topology of the network either, but in the individual users journeying through the network. These two forces, one intrinsic to the task of search engines themselves, the other intrinsic to what has become their dominant business model – advertising – drive the creation of the second index. This one is not about the world's information, but about the world's users of information. While the first one is based on publicly available information created by third parties, the second one is based on proprietary information created by the search engines themselves. By overlaying the two indexes the search engines hope to improve their core tasks: to deliver relevant search results to the users, and to deliver relevant users to advertisers. In the process, search engines create a new model of how to organize the world's information, a model composed of two self-referential worlds: one operating on the collective level – one world emerging from the interactions of everyone (at least in its ideal version) – and the other operating on the individual level – one's world as emerging from one's individual history. Both of these levels are highly dynamic and by interrelating them, search engines aim to overcome the problem of information overload (too much irrelevant information), which both users and customers constantly encounter, the former when being presented with hundreds of "hits" while looking for just a handful (maybe even just one), the latter when having to relate to a mass of people who don't care, rather than just a few prospective customers. Speaking about the user's problem, Peter Fleischer, Google's global privacy counsel, writes:

> Developing more personalized search results is crucial given how much new data is coming online every day. The University of California Berkeley estimates that humankind created five exabytes of information in 2002 – double the amount generated in 1999. An exabyte is a one followed by 18 noughts. In a world of unlimited information and limited time, more targeted

> and personal results can really add to people's quality of life.[4]

Fleischer and others usually present this as a straightforward optimization drive – improved search results and improved advertising. While this is certainly part of the story, it is not everything. This development also raises a number of troubling issues, ranging from surveillance, understood both as individual tracking and social sorting, to a potentially profound loss of autonomy. The latter is related to the fact that we are presented with a picture of the world (at least how it appears in search results) made up of what someone else, based on proprietary knowledge, determines to be suitable to one's individual subjectivity. In the following, we will try to address these issues. First, we will provide a sense of the scale of this second index as it is being compiled by Google, the most prolific gatherer of such data.

Given the sensitive and proprietary nature of that index, we can only list some publicly visible means through which this information is generated and look at the particular types of information thus gathered. We have no inside knowledge about what is being done with that information. Based on this overview, we will discuss some aspects of surveillance and of personalization. It is key to note that these are not issues we can reasonably opt out of. The pressures that are working to create and use the second index are real, and they are also driven by legitimate user needs. Thus, the question raised at the end of this text is how to address these challenges in an adequately nuanced way.

Gathering Data

Since its inception, Google has been accumulating enormous amounts of information about its users. Yet surveys have shown that most users are not aware of the wide range of data collected; many still have an understanding of Google as a search engine rather than a multi-million dollar corporation making large profits from devising personalized advertising schemes.[5] While we concentrate on Google, it is important to note that it does not differ from its competitors in principle. Other search engines like Yahoo! and Live also enable cookies to track users' search histories; stores like Amazon.com also store information on their customers' shopping habits or credit card numbers; social networking sites like Facebook have had their own share of privacy problems, such as when users found out they could not permanently delete their accounts.[6] Cloud computing is promoted by companies like Microsoft or Dell, too. The case of Google, however, is special because of the sheer quantity of data that is gathered, because of the enormous variety and quality of this information, and because of Google's growing market dominance in the fields of web search, advertising, and cloud computing.

Google employs manifold methods to collect data on its users. Click tracking has long enabled Google to log any click made by anyone on any of its servers. Log files store any server request ever made to any of Google's servers and always save basic details like the user's IP address, location, date, time, time zone, language, operating system, and browser ("standard log information"). Additional information is also sent back to Google and logged thanks to JavaScript and Web Beacons being embedded in websites.[7] Cookies are used on all Google sites. Originally, these cookies were set to expire in 2038. In 2007 the system was changed so that currently Google cookies have a life expectancy of two years – unless you use a Google service in that time, which prompts a renewal by another two years, thereby rendering the cookie virtually immortal. Cookies reporting back to Google are not only used on sites that are on Google's web properties, but also on seemingly unrelated sites which use AdSense, AdWords, or DoubleClick ads or the statistics tool Google Analytics. Most users, and even many web-administrators, are not aware of technical implications of those services, i.e. that these cookies help to collect data on the users' movement around a large percentage of existing off-Google websites. In addition to these fairly opaque techniques, which are known only to Google's more tech savvy audience, there are also data sets provided by the users themselves. Users voluntarily fill in forms in order to sign up for accounts, which results in a log of those very personal details. It was only by employing this function and by combining information collected from its wide range of services and from third parties that Google could begin to accumulate the expansive knowledge about its users that has made it so powerful and attractive to advertising partners. For purposes of analysis we can differentiate between three types of profiles, which together create a comprehensive profile of each user. First, users are tracked as "knowledge beings" and, in a second step, as "social beings", exploiting their real identities, contacts, and interaction with those contacts. A final data set captures the users as physical beings in real space. The following survey is by no means a comprehensive collection of Google's services, not least of all because they are constantly expanding. Yet we believe that the services which will be added in the future will further contribute to increasing the scope of any (or all) of these three profiles.

Layer 1: Assembling a Knowledge Profile

The basic service of Google Search has never required its users to register for an account or to actively provide any personal details. Standard log information (see above) is, however, always gathered in the background. In addition, Google servers record information specific to the search function: the search query itself, default language settings, the country code top-level domain (whether google. com or google.at is being used), the search engine result pages and the number of search results, as well as settings like safe search (filtering and blocking adult content), or additional search options like file type, and region.

Various other services can arguably be placed in the same category, firstly because they let users search for web information provided by third parties – be it websites, magazines, books, pictures, or products – and secondly because Google employs them to gather similar sets of data about its users. These services include Google Directory, Image Search, News Search, News Archive Search, Google Scholar, Google Books, Google Video, Blog Search, the LIFE photo archive hosted by Google, or Google Product Search (the service formerly known as Froogle).

Hal Roberts from the Berkman Center for Internet & Society at Harvard University writes that,

> [i]n fact, it is likely that this collection of search terms, IP addresses, and cookies represents perhaps the largest, most sensitive single collection of data extant, on- or offline. Google may or may not choose to do the relatively easy work necessary to translate its collection of search data into a database of personally identifiable data, but it does have the data and the ability to query personal data out of the collection at any time if it chooses (or is made to choose by a government, intruder, disgruntled worker, etc).[8]

The conception of services like AdSense, AdWords, AdPlanner, or Analytics[9] has helped Google to spread even further: Google Analytics can be installed on any website for free, but in exchange Google is entitled to collect information from all sites using its tool. Some estimates say that "the software is integrated into 80 percent of frequently visited German-language Internet sites".[10] The same principle governs Google's advertising schemes. Visitors to off-Google sites are usually not aware that Google servers could nevertheless be logging their online activities because "[w]eb companies once could monitor the actions of consumers only on their own sites. But over the last couple of years, the Internet giants have spread their reach by acting as intermediaries that place ads on thousands of Web sites, and now can follow people's activities on far more sites."[11]

Google's new web browser Chrome, a good example of the development towards the Google Cloud, must also be seen as a tool to construct the user's knowledge profile – with the additional benefit that it blurs the distinction of what the user "knows" online and offline. "Chrome's design bridges the gap between desktop and so-called 'cloud computing'. At the touch of a button, Chrome lets you make a desktop, Start menu, or Quick Launch shortcut to any Web page or Web application, blurring the line between what's online and what's inside your PC."[12] Chrome's "Omnibox" feature, an address bar which works like a Google search window with an auto-suggest function, logs every character typed by its users even without hitting the enter button.[13] Faced with severe criticism, Google maintained that the data collected by this keystroke logging process (keystrokes and associated data such as the user's IP address) would be rendered anonymous within 24 hours.[14] Concerns have also been voiced regarding Chrome's history search feature, which indexes and stores sensitive user data like personal financial or medical information even on secure (https://) pages.[15]

New services are introduced to systematically cover data which has not previously been included in the user profile. In the process, basic personal data is enriched by and may be combined with information covering virtually each and every aspect of a user's information needs: Google servers monitor and log what users search for (Google Search, Google Toolbar, Google Web History), which websites they visit (Google Chrome, Web Accelerator), which files (documents, e-mails, chat, web history) they have on their computers (Google Desktop), what they read and bookmark (Google Books incl. personalized library, Google Notebook, Google Bookmarks, Google Reader), what they write about in e-mails (Gmail), in work files and personal documents online (Google Docs) and offline (Google Desktop), in discussion groups (Google Groups), blogs (Blogger), or chats (Google Talk), what they want translated and which languages are involved (Google Translate).

Layer 2: Constructing the User as a Social Being

With the advent of Web 2.0 it became clear that users are not just knowledge-seeking individuals, but also social beings intensely connected to each other online and offline. Google's services have been expanding and more personalized programs have been added to its portfolio to capture the social persona and its connections.

Signing up for a Google Account allows users to access a wide range of services, most of which are not available "outside" the account. If you choose to edit your profile, basic personal details gleaned by Google from your account are not only the number of logins, your password, and e-mail address, but also your location, real-life first name and last name, nickname, address, additional e-mail address(es), phone number(s), date and place of birth, places you've lived in, profession, occupations past and present, a short biography, interests, photos, and custom links to online photo albums, social network profiles, and personal websites.

In addition, Google acquired previously independent sites like YouTube and Blogger, "migrated" their account information, and launched additional social networks like Orkut and Lively, Chat, Calendar, mobile phones and other social services.

On top of monitoring all the contents of communications (e-mails, postings, text messages sent or received), services like Google Groups, Gmail, Google Talk, Friends Connect, or Orkut store any external text, images, photos, videos, and audio files submitted, contact lists, groups a user joins or manages, messages or topics tracked, custom pages users create or edit, and ratings they make. Google servers also record who users network with (Orkut), where, when, and why they meet friends and work contacts, which friends replied to invitations or what these contacts' e-mail addresses are (Google Calendar), where they do their online shopping and which products they look for (Catalog Search, Product Search, Google Store), which credit cards they use and what their card expiration date(s) and card verification number(s) are, where they buy from, where they want their purchase shipped, how much they bought and at what price, who they bought from, and which type of payment was used (Google Checkout, Google Video). In addition, stock portfolio information, i.e. stocks selected, the amount of a user's shares, and the date, time, and price at which they were bought, is collected as well (Google Finance).

All of these services transmit information, such as account activity, passwords, login time and frequency, location, size and frequency of data transfers, preferred settings, and all clicks, including UI elements, ads, and links. All this information is subsequently saved and stored in Google's log files. This monitoring of the users' social interactions gives Google the means to create ever more personalized data profiles. It is important to note that Google originally denied intentions that it was already connecting or planning to connect information from its various services.[16] In 2004, however, the company was forced to make its privacy policy more understandable due to a new Californian law[17] and the new wording included a passage stating that Google was allowed to "share the information submitted under [any Google] account among all of [their] services in order to provide [users] with a seamless experience".

Layer 3: Recreating the User's Real-Life Embodiment

For efficient advertising, the process of constructing the user increasingly tries to capture information about the individual as an embodied person and as an agent within a real-life environment. In order to add such data to its burgeoning second index, Google collects such disparate data as its users' blood type, weight, height, allergies, and immunizations, their complete medical history and records such as lists of doctors appointments, conditions, prescriptions, procedures, and test results (Google Health). Google already knows where its users live (Google Account, Google Checkout, Google Maps default location) and where they want to go (Google Maps), which places they like (annotated maps with photos of and comments about "favorite places" in Google Maps), what their house looks like (Google Maps satellite function, Google Street View), which videos they watch (Google Video, YouTube), what their phone numbers are (Google Checkout), which mobile carrier and which mobile phones they use, and where (Dodgeball, GrandCentral, G1/ Android, MyLocation).

New data-rich sources are accessed through mobile phones, which store their owner's browsing history and sensitive personal data that identifies them, and combine these with new functions such as location tracking. Android, Google's new mobile phone platform, offers applications like Google Latitude and MyLocation, which are used to determine a user's position; MyTracks is used to track users over time.[18] Google will thereby be able to assemble even closer consumer profiles; by targeting users with geo-located ads, Google is creating a virtual gold mine for associated advertisers.[19] This foray into the mobile market enables Google to collect mobile-specific information, which can later be associated with a Google Account or with some other similar account ID. Google Latitude also offers an option to invite a "friends" list to track the inviter. Privacy groups have drawn attention to the problem that user-specific location data can also be accessed by third parties without a user's knowledge or consent and that the victim may remain unaware of being tracked[20], thereby rendering ordinary mobile phones useful tools for personal surveillance.[21]

The development of Android is perhaps the clearest example for the amount of resources Google is willing to invest in order to generate and get access to data for expanding its second index; an entire communication infrastructure is being created that is optimized for gathering data and for delivering personalized services, which are yet another way of gathering more data. Thus, Google systematically collects data to profile people in terms of their knowledge interests, social interaction, and physical embodiment. Outsiders can only observe the mechanism of gathering this data, which allows conclusions about the scope of the second index. What we cannot know is what precisely is being done with it, and even less what might be done with it in the future. What we can assess, though, are some issues arising from the mere existence of this index and the published intentions for its use.

Surveillance and Personalization

Considering the scope and detail of personal information gathered by Google (and other search engines), it is clear that very significant surveillance capacities are being generated and privacy concerns have been voiced frequently.[22] As we are trying to unravel the potential of that surveillance, it is important to distinguish three kinds of surveillance. While they are all relevant to search engines, they are actually structured very differently and subject to very different influences. First is surveillance in the classic sense of Orwell's Big Brother, carried out by a central entity with direct power over the person under surveillance, for example the state or the employer. Second is surveillance in the sense of an increasing capacity of various social actors to monitor each other, based on a distributed "surveillant assemblage" where the roles of the watchers and the watched are constantly shifting.[23] Finally, there is surveillance in the sense of social sorting, i.e. the coding of personal data into categories in order to apply differential treatment to individuals or groups.[24]

In the first case, powerful institutions use surveillance to unfairly increase and exercise their power. Search engines, however, have no direct power over their users which they can abuse. Assuming that search engines are concerned with their reputation and do not provide data from their second index to third parties, their surveillance capacities of this type appear to be relatively benign. Of course, this assumption is deeply questionable. Complex information processing systems are always vulnerable to human and technical mistakes. Large data sets are routinely lost, accidentally disclosed, and illegally accessed. Just how sensitive such search engine data can be was demonstrated in August 2006, when AOL made available the search histories of more than 650,000 persons. Even though the search histories were anonymized, the highly personal nature of the data made it possible to track down some of the users whose data had been published.[25] We can assume that large search engine companies have extensive policies and procedures in place to protect their data, thereby decreasing the likelihood of accidents and breaches. However, the massive centralization of such data makes any such incident all the more consequential, and the pull exerted by the mere existence of such large surveillance capacities is more problematic. Given the insatiable demands of business rivals, law enforcement and national security agencies, requests to access these data will certainly be made.[26] Indeed, they have already been made. In August 2005, all major search engines in the US received a subpoena to hand over records on millions of their users' queries in the context of a review of an online pornography law. Among these, Google was the only one who refused to grant access to these records.[27] While this was the only publicized incident of this kind, it is highly unlikely that this was the only such request. The expanded powers of security agencies and the vast data sets held by search engines are bound to come in close contact with one another. Historically, the collaboration between telecom companies and national security agencies has been close and extensive. Massive surveillance programs, such as ECHELON, would have been impossible without the willing support of private industry.[28] It is thus well possible that search engines are already helping to expand big-brother surveillance capacities of the state, particularly the US government.

Another more subtle effect of surveillance was first utilized by Jeremy Bentham in the design of a correctional facility (1785) and later theorized as a more general governmental technique by Michel Foucault. In a Panopticon, a place where everything can be seen by an unseen authority, the expectation of being watched affects the behavior of the watched. It is normalized, that is, behavioral patterns which are assumed to be "normal" are reinforced simply by knowing that the authority could detect and react to abnormal behavior.[29] This has also been shown to occur in the context of CCTV surveillance[30]. In the case of search engines, the situation is somewhat different because the watching is not done by a central organization; rather, the search engines provide the means for everyone to place themselves at the center and watch without being watched (at least not by the person being watched). It is not far-fetched to assume that the blurring of private and public spheres on the web impacts on a person's behavior. It is much too early to say anything systematic about the effects of this development, not least because this surveillance does not try to impose a uniform normative standard. Thus, it is not clear what normalization means under such circumstances. However, this is not related to the surveillance carried out by search engines themselves, but rather to their function of making any information accessible to the public at large and the willingness of people to publish personal information.

What is specific to the second index is the last dimension of surveillance, social sorting. David Lyon develops the concept in the following way:

> Codes, usually processed by computers, sort out transactions, interactions, visits, calls and other activities. They are invisible doors that permit access to, or exclude from participation in a myriad of events, experiences and processes. The resulting classifications are designed to influence and manage populations and persons thus directly and indirectly affecting the choices

> and chances of data subjects.[31]

Rather than treating everyone the same, social sorting allows matching people with groups to whom particular procedures, enabling, disabling or modifying behavior, are assigned. With search engines, we encounter this as personalization. It is important to note that, as David Lyon stresses, "surveillance is not itself sinister any more than discrimination is itself damaging."[32] Indeed, the basic purpose of personalization is to help search engines to improve the quality of search results. It enables them to rank results in relation to individual user preferences, rather than to network topology, and helps to disambiguate search terms based on the previous path of a person through the information landscape. Personalization of search is part of a larger trend in the informational economy towards "mass individualization", where each consumer/user is given the impression, rightly or wrongly, of being treated as a unique person within systems of production still relying on economies of scale.

Technically, this is a very difficult task for which vast amounts of personal data are needed, which, as we have seen, is being collected in a comprehensive, and systematic way. A distinction is often made between data that describes an individual on the one hand, and data which is "anonymized" and aggregated into groups on the basis of some set of common characteristics deemed relevant for some reason. This distinction plays an important role in the public debate, for example in Google's announcements in September 2008 to strengthen user privacy by "deleting" (i.e. anonymizing) user data after nine rather than 18 months.[33] Questions have also been raised about how easily this "anonymization" can be reversed by Google itself or by third parties.[34] In some cases, Google offers the option to disable cookies or use Chrome's Incognito mode ("porn mode" in colloquial usage), but instructions how to do this are well hidden, difficult to follow, and arguably only relevant for highly technologically literate users. Moreover, the US-American consumer group Consumer Watchdog has pointed out that "Chrome's Incognito mode does not confer the privacy that the mode's name suggests", as it does not actually hide the user's identity.[35] It is important to note that even if Google followed effective "anonymizing" procedures, this would matter only in terms of surveillance understood as personal tracking. If we understand it as social sorting, this has nearly no impact. The capacity to build a near infinite number of "anonymized" groups from this database and to connect individuals to these small groups for predictive purposes re-integrates anonymized and personalized data in practice. If the groups are fine-grained, all that is necessary is to match an individual to a group in order for social sorting to become effective. Thus, Hier concludes that "it is not the personal identity of the embodied individual but rather the actuarial or categorical profile of the collective which is of foremost concern."[36] In this sense, Google's claim to increase privacy is seriously misleading.

Like virtually all aspects of the growing power of search engines, personalization is deeply ambiguous in its social effects. On the one hand, it promises to offer improvements in terms of search quality, further empowering users by making accessible the information they need. This is not a small feat. By lessening the dependence on the overall network topology, personalization might also help to address one of the most frequently voiced criticisms of the dominant ranking schemes, namely that they promote popular content and thus reinforce already dominant opinions at the expense of marginal ones.[37] Instead of only relying on what the majority of other people find important, search engines can balance that with the knowledge of the idiosyncratic interest of each user (group), thus selectively elevating sources that might be obscure to the general audience, but are important to this particular user (set). The result is better access to marginal sources for people with an assumed interest in that subject area.

So far so good. But where is the boundary between supporting a particular special interest and deliberately shaping a person's behavior by presenting him or her with a view shaped by criteria not his or her own? As with social sorting procedures in general, the question here is also whether personalization increases or decreases personal autonomy. Legal scholar Frank Pasquale frames the issue in the following way:

> Meaningful autonomy requires more than simple absence of external constraint once an individual makes a choice and sets out to act upon it. At a minimum, autonomy requires a meaningful variety of choices, information of the relevant state of the world and of these alternatives, the capacity to evaluate this information and the ability to make a choice. If A controls the window through which B sees the world—if he systematically exercises power over the relevant information about the world and available alternatives and options that reaches B—then the autonomy of B is diminished. To control one's informational flows in ways that shape and constrain her choice is to limit her autonomy, whether that person is deceived or not.[38]

Thus, even in the best of worlds, personalization enhances and diminishes the autonomy of the individual user at the same time. It enhances it because it makes information available that would otherwise be harder to locate. It improves, so it is claimed, the quality of the search experience. It diminishes it because it subtly locks the users into a path-dependency that cannot adequately reflect their personal life story, but reinforces those aspects that the search engines are capable of capturing, interpreted through assumptions built into the personalizing algorithms. The differences between the personalized and the non-personalized version of the search results, as Google reiterates, are initially subtle, but likely to increase over time. A second layer of intransparency, that of the personalization algorithms, is placed on top of the intransparency of the general search algorithms. Of course, it is always possible to opt out of personalization by simply signing out of the account. But the ever increasing range of services offered through a uniform log-in actively works against this option. As in other areas, protecting one's privacy is rendered burdensome and thus something few people are actively engaged in, especially given the lack of direct negative consequences for not doing it.

And we hardly live in the best of all worlds. The primary loyalty of search engines is – it needs to be noted again – not to users but to advertisers. Of course, search engines need to attract and retain users in order to be attractive advertisers, but the case of commercial TV provides ample evidence that this does not mean that user interests are always foregrounded.

The boundary between actively supporting individual users in their particular search history and manipulating users by presenting them with an intentionally biased set of results is blurry, not least because we actually want search engines to be biased and make sharp distinctions between relevant and irrelevant information. There are two main problems with personalization in this regard. On the one hand, personalization algorithms will have a limited grasp of our lives. Only selective aspects of our behavior are collected (those that leave traces in accessible places), and the algorithms will apply their own interpretations to this data, based on the dominant world-view, technical capacities and the particular goals pursued by the companies that are implementing them. On the other hand, personalization renders search engines practically immune to systematic, critical evaluation because it is becoming unclear whether the (dis)appearance of a source is a feature (personalization done right) or a bug (censorship or manipulation).

Comparing results among users and over time will do little to show how search engines tweak their ranking technology, since each user will have different results. This will exacerbate the problem already present at the moment: that it is impossible to tell whether the ranking of results changes due to aggregate changes in the network topology, or due to changes in the ranking algorithms, and, should it be the latter, whether these changes are merely improvements to the quality, or attempts to punish unruly behavior.[39] In the end, it all boils down to trust and expediency. The problem with trust is that, given the essential opacity of the ranking and personalization algorithms, there is little basis to evaluate such trust. But at least it is something that is collectively generated. Thus, a breach of this trust, even if it affects only a small group of users, will decrease the trust of everyone towards the particular service. Expediency, on the other hand, is a practical measure. Everyone can decide for themselves whether the search results delivered are relevant to them. Yet, in an environment of information overload, even the most censored search will bring up more search results that anyone can handle. Thus, unless the user has prior knowledge of the subject area, she cannot know what is not included. Thus, search results always seem amazingly relevant, even if they leave out a lot of relevant material.[40]

With personalization, we enter uncharted territory in its enabling and constraining dimensions. We simply do not know whether this will lead to a greater variety of information providers entering the realm of visibility, or whether it will subtly but profoundly shape the world we can see, if search engines promote a filtering agenda other than our own. Given the extreme power differential between individual users and the search engines, there is no question who will be in a better position to advance their agenda at the expense of the other party.

The reactions to and debates around these three different kinds of surveillance are likely to be very different. As David Lyon remarked,

> Paradoxically, then, the sharp end of the panoptic spectrum may generate moments of refusal and resistance that militate against the production of docile bodies, whereas the soft end seems to seduce participants into stunning conformity of which some seem scarcely conscious.[41]

In our context, state surveillance facilitated by search engine data belongs to the sharp end, and personalization to the soft end where surveillance and care, the disabling and enabling of social sorting are hard to distinguish and thus open the door to subtle manipulation.

Conclusion

Search engines have embarked on an extremely ambitious project, organizing the world's information. As we are moving deeper into dynamic, informationrich environments, their importance is set to increase. In order to overcome the limitations of a general topological organization of information and to develop a personalized framework, a new model of the world is assumed: everyone's world is different. To implement this, search engines need to know an almost infinite amount of information about their users. As we have seen, data is gathered systematically to profile individuals on three interrelated levels – as a knowledge person, a social person, and as an embodied person. New services are developed with an eye on adding even more data to these profiles and filling in remaining gaps. Taken together, all this information creates what we call the second index, a closed, proprietary set of data about the world's users of information. While we acknowledge the potential of personalized services, which is hard to overlook considering how busy the service providers are in touting the golden future of perfect search, we think it is necessary to highlight four deeply troublesome dimensions of this second index. First, simply storing such vast amounts of data creates demands to access and use it, both within and outside the institutions. Some of these demands are made publicly through the court system, some of them are made secretly. Of course this is not an issue particular to search engines, but one which concerns all organizations that process and store vast amounts of personal information and the piecemeal privacy and data protection laws that govern all of them. It is clear that in order to deal with this expansion of surveillance technology / infrastructure, we will need to strengthen privacy and data protection laws and the means of parliamentary control over the executive organs of the state. Second, personalization is inherently based upon a distorted profile of the individual user. A search engine can never "know" a person in a social sense of knowing. It can only compile data that can be easily captured through its particular methods. Thus, rather than being able to draw a comprehensive picture of the person, the search engine's picture is overly detailed in some aspects and extremely incomplete in others. This is particularly problematic given the fact that this second index is compiled to serve the advertisers' interests at least as much as the users'. Any conclusion derived from this incomplete picture must be partially incorrect, thus potentially reinforcing those sets of behaviors that lend themselves to data capturing and discouraging others. Third, the problem of the algorithmic bias in the capturing of data is exacerbated by the bias in the interpretation of this data. The mistakes range from the obvious to the subtle. For example, in December 2007 Google Reader released a new feature, which automatically linked all one's shared items from Reader with one's contact list from Gmail's chat feature, Google Talk. As user banzaimonkey stated in his post, "I think the basic mistake here [...] is that the people on my contact list are not necessarily my 'friends'. I have business contacts, school contacts, family contacts, etc., and not only do I not really have any interest in seeing all of their feed information, I don't want them seeing mine either."[42] Thus, the simplistic interpretation of the data – all contacts are friends and all friends are equal – turned out to be patently wrong. Indeed, the very assumption that data is easy to interpret can be deeply misleading, even the most basic assumption that users search for information that they are personally interested in (rather than someone else). We often do things for other people. The social boundaries between persons are not so clear cut.

Finally, personalization further skews the balance of power between the search engines and the individual user. If future search engines operate under the assumption that everyone's world is different, this will effectively make it more difficult to create shared experiences, particularly as they relate to the search engine itself. We will all be forced to trust a system that is becoming even more opaque to us, and we will have to do this based on the criterion of expediency of search results which, in a context of information overload, is deeply misleading. While it would be as short-sighted as unrealistic to forgo the potential of personalization, it seems necessary to broaden the basis on which we can trust these services. Personalized, individualized experience is not enough. We need collective means of oversight, and some of them will need to be regulatory. In the past, legal and social pressure were necessary to force companies to grant consumers basic rights. Google is no exception: currently, its privacy policies are not transparent and do not restrict Google in substantial ways. The fact that the company's various privacy policies are hard to locate and that it is extremely difficult to retrace their development over time, cannot be an oversight – especially since Google's mission is ostensibly to organize the world's information and to make it more easily accessible to its users. But regulation and public outrage can only react to what is happening, which is difficult given the dynamism of this area. Thus, we also need means oriented towards open source processes, where actors with heterogeneous value sets have the means to fully evaluate the capacities of a system. Blind trust and the nice sounding slogan "don't be evil" are not enough to ensure that our liberty and autonomy are safeguarded.

Notes

[1] All links were last accessed on 25 Mar 2009.

[2] Shirky, Clay, "Ontology is Overrated: Categories, Links, and Tags", 2005
http://shirky.com/writings/ontology_overrated.html

[3] Demoz.org. http://www.demoz.org

[4] Fleischer, Peter, Google's search policy puts the user in charge. FT.com, May 25, 2007.
http://www.ft.com/cms/s/2/560c6a06-0a63-11dc-93ae-000b5df10621.html

[5] Electronic Privacy Information Center. "Search Engine Privacy", 3 Feb 2009.
http://epic.org/privacy/search_engine

[6] Aspan, Maria, "How Sticky Is Membership on Facebook? Just Try Breaking Free", New York Times Online, 11 Feb 2008. http://www.nytimes.com/2008/02/11/technology/11facebook.html

[7] Dover, Danny, The Evil Side of Google? Exploring Google's User Data Collection, SEOmoz.org, 24 Jun 2008. http://www.seomoz.org/blog/the-evil-side-of-google-exploring-googles-use...

[8] Roberts, Hal, "Google Privacy Videos", 6 March 2008. http://blogs.law.harvard.edu/hroberts/2008/03/06/google-watching-persona...

[9] Bogatin, Donna, "Google Analyitcs: Should Google be minding YOUR Web business?", ZDNet Blogs, 9 May 2007. http://blogs.zdnet.com/micro-markets/?p=1324

[10] Bonstein, Julia, Marcel Rosenbach and Hilmar Schmundt, "Data Mining You to Death: Does Google Know Too Much?" Spiegel Online International, 30 Oct 2008. http://www.spiegel.de/international/germany/0,1518,587546,00.html

[11] Story, Louise, "To Aim Ads, Web Is Keeping Closer Eye on You", New York Times.com, 10 Mar 2008. http://www.nytimes.com/2008/03/10/technology/10privacy.html?_r=1

[12] Mediati, Nick, "Google's streamlined and speedy browser offers strong integrated search and an intriguing alternative to Firefox and Internet Explorer", PCWorld.com, 12 Dec 2008. http://www.pcworld.com/article/150579/google_chrome_web_browser.html

[13] Fried, Ina, "Google's Omnibox could be Pandora's box", CNet.com, 3 Sep 2008. CNet.com http://news.cnet.com/8301-13860_3-10031661-56.html

[14] Cheung, Humphrey, "Chrome is a security nightmare, indexes your bank accounts", TGDaily.com, 4 Sep 2008. http://www.tgdaily.com/content/view/39176/108

[15] ibid.

[16] Electronic Frontier Foundation, "Google's Gmail and Your Privacy –What's the Deal?" EFFector Vol.17, No.12, 09 Apr 2004. http://w2.eff.org/effector/17/12.php#I; Liedtke, Michael, "Consumer watchdogs tear into Google's new e-mail service", USAToday.com, 07 Apr 2004. http://www.usatoday.com/tech/news/internetprivacy/2004-04-07-gmail-bad-k... Hafner, Katie, "In Google We Trust? When the Subject is E-Mail, Maybe Not", New York Times online, 08 Apr 2004. http://www.nytimes.com/2004/04/08/technology/circuits/08goog.html?ei=500...

[17] Olsen, Stefanie, "California privacy law kicks in", CNet.com, 06 July 2004. http://news.cnet.com/California-privacy-law-kicks-in/2100-1028_3-5258824...

[18] Hodgin, Rick C., "Google launches My Tracks for Android", TGDaily.com, 13 Feb 2009. http://www.tgdaily.com/content/view/41439/140/

[19] Greenberg, Andy, "Privacy Groups Target Android, Mobile Marketers", Forbes.com, 13 Jan 2009. http://www.forbes.com/2009/01/12/mobile-marketing-privacy-tech-security-...

[20] Privacy International.Org, "Privacy international identifies major security flaw in Google's global phone tracking system", 05 Feb 2009. http://www.privacyinternational.org/article.shtml?cmd[347]=x-347-563567

[21] Electronic Privacy Information Center: "Personal Surveillance Technologies", 24 Nov 2008.http://epic.org/privacy/dv/personal_surveillance.html

[22] For a summary, see Alexander Havalais, Search Engine Society (Cambridge UK: Polity Press) 2009, 139-159

[23] Kevin D, Haggerty and Richard V. Ericson, "The Surveillant Assemblage", British Journal of Sociology, December 2000, Vol. 51 No. 4, pp. 605–622

[24] See, Gandy, Jr., Oscar H., The Panoptic Sort. A Political Economy of Personal Information (Boulder: Westview Press), 1993; Lyon, David (ed.), Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination, (London, New York: Routledge) 2002

[25]"A Face Is Exposed for AOL Searcher No. 4417749", NYT.com (August 9 2006)

[26]Jonathan Zittrain, professor of Internet law at Harvard, explains: "This is a broader truth about the law. There are often no requirements to keep records, but if they're kept, they're fair game for a subpoena." Quoted in: Noam Cohen, "As data collecting grows, privacy erodes", International Herald Tribune (February 16 2009.)

[27]Hafner, Katie, Matt Richtel, "Google Resists U.S. Subpoena of Search Data", NYT.com (January 20 2006.)

[28]Final Report on the existence of a global system for the interception of private and commercial communications (ECHELON interception system), European Parliament Temporary Committee on the ECHELON Interception System, approved September 5 2001

[29]Vaz, Paulo, and Fernanda Bruno, "Types of Self-Surveillance: from abnormality to individuals 'at risk'", Surveillance & Society, Vol. 1, No. 3, 2003.

[30]Norris, Clive, Gary Armstrong, The Maximum Surveillance Society: The Rise of CCTV, Oxford, UK, Berg, 1999

[31]Lyon, David, 2002, p.13

[32]Ibid.

[33]Fleischer, Peter, "Another step to protect user privacy", Official Google Blog, 8.Sep 2008, http://googleblog.blogspot.com/2008/09/another-step-to-protect-user-priv...

[34]Soghoian, Chris, "Debunking Google's log anonymization propaganda", Cnet.com, 11 Sep, 2008. http://news.cnet.com/8301-13739_3-10038963-46.html

[35]Find videos at http://www.consumerwatchdog.org/corporateering/corpact4/

[36]Hier, Sean P., "Probing the Surveillant Assemblage. On the dialectics of surveillance practices as processes of social control", Surveillance & Society, Vol. 1: 3, 2003, 399–411

[37]Among the first to voice this critique was Lucas Introna and Helen Nissenbaum, "Shaping The Web: Why The Politics of Search Engines Matters," Information Society 16, No. 3 (2000): 169 185

[38]Pasquale, Frank A.; Bracha, Oren (2007), "Federal Search Commission?: Access, Fairness and Accountability in the Law of Search",.. Public Law and Legal Theory Research Paper No. 123 July: Social Science Research Network at http://ssrn.com/abstract=1002453

[39]Greenberg, Andy, "Condemned To Google Hell", Forbes.com (April 30, 2007)

[40]Bing Pan et.al., "In Google We Trust: Users' Decisions on Rank, Position, and Relevance", Journal of Computer Mediated Communication, Vol.12, 3, 2007

[41]Lyon, David, "Introduction", In: David Lyon (ed.) Theorizing Surveillance. The Panopticon and Beyond, Devon, UK, Willian Publishing, 2006, 8

[42]Google Reader Help, "New Feature: Sharing with Friends", 14 Dec 2007 ff. http://groups.google.com/group/google-reader-howdoi/browse_thread/thread...

Stalder, Felix. Mayer, Christine (2009). The Second Index. Search Engines, Personalization and Surveillance. In: Konrad Becker/Felix Stalder (eds.). Deep Search. The Politics of Search beyond Google. Studienverlag. Distributed by Transaction Publishers, New Jersey

10 February, 2010 - 06:00 in

- my_publications
- search

- felix's blog
- Add new comment

- 🖨
- ✉

Search

Notes & nodes on society, technology and the space of the possible, by Felix Stalder.

Deep Search II, Conference, Vienna, May 28 2010

my books

- Deep Search. The Politics of Search Beyond Google Studienverlag/Transaction Publishers, 2009

- Media Arts Zurich. 13 Positions / Mediale Kunst Zürich. 13 Positionen. Scheidegger & Spiess 2008.
- Manuel Castells and the Theory of the Network Society. Polity Press, 2006
- Open Cultures and the Nature of Networks. edited by Kuda.org, Futura publikacije, Novi Sad and Revolver - Archiv für aktuelle Kunst, Frankfurt a.M. (October 2005)

my recent articles

- Autonomy and Control in the Era of Post-Privacy
- Das Urheberrecht ist alt geworden. 300 Jahre alt.
- Digital Commons
- Kritische Strategien zu Kunst und Urheberrecht
- The Second Index. Search Engines, Personalization and Surveillance (Deep Search)
- Deep Search: Introduction
- Nachahmung, Transformation und Autorfunktion
- Deep Search: The Politics of Search Beyond Google
- Neun Thesen zur Remix-Kultur
- Analysis Without Analysis (Clay Shirky Review)
- Bourgeois anarchism and authoritarian democracies (First Monday, 07.2008)
- 30 Years of Tactical Media (book chapter)
- Torrents of Desire and the Shape of the Information Landscape (book chapter)
- On the Differences between Open Source and Open Culture (book chapter)
- Neue Formen der Öffentlichkeit und kulturellen Innovation zwischen Copyleft, Creative Commons und Public Domain. (Buchkapitel)

more

their recent tweets

- Autonomy & Control Post-Privacy: Felix Stalder http://j.mp/baDpyb Networked individuals vs personalised institutions /via @josswinn
- Stichting Kunst en Openbare Ruimte - Felix Stalder, Autonomy and Control in the Era of Post-Privacy http://ff.im/-n50OX
- Stichting Kunst en Openbare Ruimte - Felix Stalder, Autonomy and Control in the Era of Post-Privacy http://bit.ly/dfzHre #privacy
- Autonomy and Control in the Era of Post-Privacy: Researcher Felix Stalder analyses the loss of the key role of the... http://bit.ly/cusyB3
- RT @fugitivephilo: Felix Stalder on post-privacy , control & autonomy #skor #cahier #autonomia http://bit.ly/duNC07

more

Navigation

- News aggregator