Shaping the Web:
Why the politics of search engines matters

Lucas D. Introna
London School of Economics
l.introna@lse.ac.uk

Helen Nissenbaum
University Center for Human Values
Princeton University
helen@Princeton.edu

ABSTRACT

This article argues that search engines raise not merely technical issues but also political ones. Our study of search engines suggests that they systematically exclude (in some cases by design and in some accidentally) certain sites, and certain types of sites, in favor of others, systematically give prominence to some at the expense of others. We argue that such biases, which would lead to a narrowing of the Web's functioning in society, run counter to the basic architecture of the Web as well as the values and ideals that have fueled widespread support for its growth and development. We consider ways of addressing the politics of search engines, raising doubts whether, in particular, the market mechanism could serve as an acceptable corrective.

INTRODUCTION

The Internet, no longer merely an e-mail and file-sharing system, has emerged as a dominant interactive medium. Enhanced by the technology of the World Wide Web, it has become an integral part of the ever-expanding global media system, moving into center stage of media politics alongside traditional broadcast media – television and radio. Enthusiasts of the "new medium" have heralded it as a democratizing force that will give voice to diverse social, economic and cultural groups, to members of society not frequently heard in the public sphere. It will

empower the traditionally disempowered, giving them access both to typically unreachable nodes of power and previously inaccessible troves of information.

To scholars of traditional media, these optimistic claims must have a ring of familiarity, echoing similar optimistic predictions concerning the democratizing and empowering capacities of both radio and television. Instead of the expected public gains and fulfilment of democratic possibilities, instead of the spreading of access and power, however, the gains, the power, and the access were consolidated in the hands of a few dominant individuals and institutions. In the words of acclaimed media critic Robert McChesney (1999),

> The American media system is spinning out of control in a hyper-commercialized frenzy. Fewer than ten transnational media conglomerates dominate much of our media; fewer than two dozen account for the overwhelming majority of our newspapers, magazines, films, television, radio, and books. With every aspect of our media culture now fair game for commercial exploitation, we can look forward to the full-scale commercialization of sports, arts, and education, the disappearance of notions of public service from public discourse, and the degeneration of journalism, political coverage, and children's programming under commercial pressure.

McChesney's work (1993, 1997) traces—in very subtle and convincing detail—how commercial interests were woven into the very fiber of the modern media networks through legislation, market mechanisms, and the like. These moves progressively pushed out and silenced the public service agenda, which was very central to the vision of the early pioneers in the field—McChesney's (1993) historical account of radio is very telling in this regard. His central argument, historically grounded, is that the fundamental course of media is determined primarily by *how they're owned and operated*. Most US communication media—going back to AM radio in the 1920s—have followed this path: At first, when they do not seem commercially viable, they are developed by the nonprofit, noncommercial sector. When their profit-making potential emerges, however, the corporate sector start colonizing the media, and through a variety of mechanisms, usually its dominance of politicians, muscle out the rest and take over. McChesney argues that this pattern is seen in the cases of FM radio, in UHF television, and to some extent in satellite and cable.

On the prospects of the Internet, there are divergent predictions. Some, like Dan Schiller

(1995) and McChesney, influenced by their knowledge of other media, anticipate a similar narrowing of prospects for the Internet. They point to the commitment of the United States to private ownership of communications technology as the single most important and consistent historical policy position that influenced the course of telecommunications development. And this same commitment is clearly evident in the rhetoric of the political foundations of the Internet, namely, the fact that of five 'values' that Vice-President Gore identified as ones that should define and guide the development of the GII, the first one listed was "private investment" (Office of the Vice President, 1995). Schiller asks, "What is the likelihood of robust adherence to…elemental democratic prescription, when the character of the network development is now all-too-evidently to be given mainly as a function of unrestrained corporate ambition and private design?" (Schiller, 1995, p.6). Others, like Mark Poster (1995), offer a contrasting view, arguing that the distinctly "post-modern" nature of the Internet, with its capacity to disseminate material rather than centralize it, will discourage the endowment of authority – both academic as well as political. Its development, therefore, is unlikely to mirror that of previous media.

The broader debate about the dual possibilities of media: to be democratizing or to be colonized by specialized interests at the expense of the public good, inspires and motivates this article on the politics of search engines. The general position we defend, and illustrate in this one case, is that although the Internet and the Web offer exciting prospects for furthering the public good, the benefits are conditional, resting precariously on a number of political, economic and technical factors. Following Poster, we are buoyed by clear instances where the Web and Internet have served broad political and ends. But we also see irrefutable signs of gradual centralization and commercialization of guiding forces. Like McChesney, we are particularly concerned with the way these competing interests (centralized commercial versus decentralized public) may, early on, be woven in, or out, of the very fiber of media networks. Search engines constitute a particularly telling venue for this competition. And prospects, as seen from the perspective of the time of writing this paper, do not look good for broad public interests.

Search engines constitute a powerful source of access and accessibility within the Web. Access, already a thorny issue, is the subject of much scholarship and research (Golding, 1994; Hoffman and Novak, 1998; Pollack, 1995; Schiller, 1995), as well as a lengthy report of the National Telecommunications and Information Administration (NTIA) *Falling through the Net*.

Focussing on social, economic and racial factors, these works show how access to the Web is preconfigured in subtle but politically important ways, resulting in exclusion of significant voices. It is not enough, however, to worry about overcoming these traditional barriers, to focus only on the granting of *entry* to the media space of the Web. It is not enough if, as we argue, the space itself is distorted in favor of those wealthy in technical or economic resources through the mechanism of biased search engines. The politics of search engines thus represents the broader struggle to sustain the democratic potential of traditional media, the Internet, and the World Wide Web in particular.

In a statistical study of Web search engines, S. Lawrence and C.L. Giles (1999) estimated that none of the search engines they studied, taken individually, index more than 16% of the total indexable Web, which they estimate to consist of 800 million pages.[1] Combining the results of the search engines they studied, they estimated the coverage to increase to approximately 42%. This confirms the primitive impressions of many users, namely, that the Web is almost inconceivably large, and also that search engines only very partially meet the desperate need for an effective way of finding things.[2] When judging what the producers of search engines have accomplished so far, optimists, focusing on the half-full portion of the cup, may legitimately marvel at progress in Web search technologies and at the sheer bulk of pages that are successfully found. In this paper, however, we are concerned with the half-empty portion of the cup: the portions of the Web that remain hidden from view.

The purpose of this article is not, however, to bemoan the general difficulties of building comprehensive search engines, nor to highlight the technological difficulties that must surely impose limits on the range of scope and coverage that even the best search engines can achieve.

---

1 S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature*, 400, 107-109, 1999.

2 In an on-line survey the NDP Group polled 22000 seekers that accessed search engines to determine their satisfaction with the search engine. Ninety six percent (96%) indicated that they were satisfied with the search results. This would seem to go against our argument. However, in another study done by researchers from British Telecom (BT), PC literate, but not regular users of the Internet, found their search results disappointing and generally "not worth the effort". Pollock, A and A. Hockley, "What's Wrong with Internet Searching" *D-Lib Magazine*, March 1997. This may indicate that one needs a fairly high level of searching skills to get what you want. We will return to this issue when we discuss the market argument for the development of search engines.

Our concern, rather, is with the ways that developers, designers and producers of search engines will direct these technological limitations, the influences that may come into play in determining any systematic inclusions and exclusions, the wide-ranging factors that dictate systematic prominence for some sites, dictating systematic invisibility for others. These, we think, are political issues.[3] They are important because what people (the "seekers") are able to find on the Web determines what the Web consists of for them. And we all—individuals and institutions alike—have a great deal at stake in what the Web consists of.

A BRIEF AND SELECTIVE TECHNICAL OVERVIEW

Although a complete discussion of the technical detail of search engines is beyond the scope of this article,[4] we highlight aspects of search engines that we consider relevant to our discussion of their politics. We briefly discuss the nature of the connection between search engines and Web pages, the process by which this relationship is established, and how this relationship affects the producers (or owners) of Web pages wishing to have their pages recognized. Web-page providers seeking recognition from search engines for their Web pages must focus on two key tasks: (a) being indexed and (b) achieving a ranking in the top 10-20 search results displayed.[5]

On Being Indexed

Having a page indexed, the essential first stage of being recognized by search engines, is extremely important. Without much exaggeration one could say that to exist is to be indexed by a search engine. If a webpage is not in the index of a search engine, a person wishing to access it must know the complete URL —webpage address—such as http://is.lse.ac.uk/lucas/cepe98.html for the CEPE'98 conference.[6] Since there is no rigid standard for producing URLs they are not

---

3 Winner, L. (1980) 'Do Artifacts Have Politics', *Daedalus*, 109 , 121-136.

4 For those interested in more detail the website http://www.searchenginewatch.com is a good place to start.

5 We are here thinking of the top 10-20 when it is a matter of actual relevancy. We will later discuss the issue of spamming.

6 One could argue that it is also possible for a webpage to be found through portal sites, which are increasingly popular, though as a matter for fact, we think it would be highly unlikely

obvious or even logical, in the way we tend to think that the addresses of our physical homes are logical.[7] Sometimes the Internet domain name structure may help, such as 'ac.uk' or 'edu' for an academic institution in the United Kingdom or United States. However, for most searches we do not have any idea of the URLs involved.[8]

This is where search engines enter the picture. They create a map of the Web by indexing webpages according to keywords, and then create enormous databases that link page content, to keywords, to URLs. When a seeker of information submits a keyword (or phrase)—presumably, one that best captures her interest—the search engine database ideally returns to the seeker a list of URLs linked to that keyword, ideally including all those that are relevant to the seeker's interest. It is important to note that search engines use the notion of a keyword (i.e. that which is indexed and hence used for searching) in a rather minimal sense. Keywords are not determined a priori by the designers of the search engines' databases, nor, explicitly, by some other authority, but rather they are "deduced" from webpages themselves in the process of indexing. In a particular webpage a keyword(s) can be any of the following:

- Actual keywords indicated by the Web page designer in an HTML metatag as follows: <meta NAME="keywords" CONTENT="list of keywords">.
- All, or some of the words appearing in the title which is indicated by the HTML <TITLE> tag as follows: <TITLE>Whatever is the title of the page</TITLE>.

---

that a link is established through a portal site that does not meet the indexing criteria for search engines.

7 We realize we have not listed all the means through which pages may be found. For example, one may access a page through an outlink from another page. The problem with such means is that they depend on somewhat unpredictable serendipity. One needs also to add that there are increasingly a variety of alternatives emerging as viable options such as portal sites and keyword retrieval via Centraal's Real Name system (http://www.centraal.com). Nevertheless, the majority of those that access the web continue to do it through search engines. There is no reason to believe that this would change in the foreseeable future.

8 We note, for readers who are aware of the debate currently raging over Domain Names, that an effective system of search and retrieval is a constructive response to the debate that would lessen the impact of whatever decisions are made. We would argue that Domain Names are important in inverse proportion to the efficacy of available search mechanism, for if individuals and institutions can easily be found on the basis of content and relevancy, there is less at stake in the precise formulation of their Domain Names. In other words, a highly effective indexing and retrieval mechanism can mitigate the effects of domain name assignments.

- The first X words in a Web page (possibly excluding stop words[9]).
- All the words in the Web page (possibly excluding stop words).

Most search engines use at least some of the words in the *title tag* of the webpage as the relevant keywords for indexing purposes.[10] It is obviously important for webpage producers as well as seekers to know what words on a particular webpage are 'seen' as keywords by the indexing software of search engines. Thus, one might naturally ask: How does a search engine go about creating its database and what does it store in it?

The answer to this question depends on which of basically two categories (and within these categories, the further subcategories) the search engine fits. One category includes directory-based search engines such as Yahoo! and Aliweb. In this category, the vast majority of the pages indexed are manually submitted to the search engines' editors by Webmasters (and other creators of Web pages).[11] The other category includes search engines that automatically harvest URLs by means of spiders (also referred to as robots or softbots). Among the most well-known search engines fitting this category are: Alta Vista, Lycos, and Hotbot.

In the case of directory-based search engines, Web-page creators submit URLs to the search engines for possible inclusion into their databases. If you wanted your page recognized by Yahoo!, for example, you would submit your URL and background information to a human editor who would review the page(s) and decide whether or not to schedule your page for indexing. If your page is scheduled for indexing, it would be retrieved by the indexing software, which would

---

9 A stop word is a frequently occurring word that is excluded because there are too many frequencies such as 'the', 'to', 'we', and so forth. Stop words are not indexed. This is not insignificant if one considers that the word 'web' is a stop word in Alta Vista. So if you are a company doing Web design that have 'Web design' in your title you may not get indexed/ranked accordingly.

10 The <TITLE> tag is either created by the webpage designer or 'deduced' by a converter. For example when you create a MSWord document and want to publish it on the Web you can save it as HTML directly in the MSWord editor. In this case the MSWord editor assumes that the first sentence it can find in the document is the tile and will place this in the <TITLE> tag in the HTML source code it generates.

11 Most of the directory based search engines also use some form of automatic harvesting to augment their manually submitted database.

parse[12] the page and index it according to the keywords (content) found in the page. For directory-based search engines, therefore, human gatekeepers hold the key to inclusion in their indexed databases. At the time of writing this article, there is a considerable backlog so that this process can take up to six months from the time of submission to the time of inclusion.

Web owners wishing to have their pages indexed must surely wonder what criteria these human editors use to decide whether or not to index their pages? This is a major bone of contention, especially for anyone contesting these decision criteria. With Yahoo!, for example, representatives say that they use criteria of relevancy (Phua, 1998).[13] The exact nature of these criteria, however, is not widely known nor publicly disseminated and, evidently, these criteria are not consistently applied by the various editors. As a result you may have your page rejected (without notification) and would not know what to do to get it accepted. Danny Sullivan, the editor of *Search Engine Watch*, believes that the base success rate for any submitted page being listed with Yahoo! is approximately 25%. Two factors that seem to increase the chances of being listed are the number of links (to and from a given site—also referred to as inlinks and outlinks), and how full a particular category happens to be. Where editors feel they need more references within a category, they lower the entry barriers. Defending their approach, representatives of Yahoo! maintain they list what users want, arguing that if users were not finding relevant information they would cease using Yahoo!. (We will return to this form of response below.) With Aliweb, a very small site when compared with its competitors, users submit supplemental information about their Web-page content and keywords, as a way to help the indexing software

---

12 When parsing the page the robot views the page in HTML format and treats it as one long string of words as explained by Altivista: "Alta Vista treats every page on the Web and every article of Usenet news as a sequence of words. A word in this context means any string of letters and digits delimited either by punctuation and other non-alphabetic characters (for example, &, %, $, /, #, _, ~), or by white space (spaces, tabs, line ends, start of document, end of document). To be a word, a string of alphanumerics does not have to be spelled correctly or be found in any dictionary. All that is required is that someone typed it as a single word in a webpage or Usenet news article. Thus, the following are words if they appear delimited in a document HAL5000, Gorbachevnik, 602e21, www, http, EasierSaidThenDone, etc. The following are all considered to be two words because the internal punctuation separates them: don't, digital.com, x-y, AT&T, 3.14159, U.S., All'sFairInLoveAndWar."

13 Phua, V. (1998) "Towards a Set of Ethical Rules for Search Engines", *MSc dissertation, LSE*.

improve the quality of its indexing and hence provide better search results. Representatives of Aliweb emphasize that they do not provide comprehensive coverage; rather, they emphasize high quality search results. Because this is a small site, it is still able to index most of its submissions. As it becomes larger, it may, like its competitors, need to establish criteria for inclusion and exclusion.

Being indexed by search engines that automatically harvest URLs is a matter of being visited by a spider (also called robot, crawler, softbot, agent, etc.). Spiders usually start crawling from a historical list of URLs, especially documents with many links elsewhere, such as server lists, "What's New" pages, and other popular sites on the web. Software robots "crawl" the Web—that is, automatically traverse the web's hypertext structure—first retrieving a document, and then recursively retrieving all documents that are referenced (linked by other URLs) in the original document. Web owners interested in having their pages indexed might wish they had access to the detail of routes that robots follow when they crawl, which sites they favor, which they visit and how often, which not, and so forth. This, however, is a complicated technical subject, which mostly is treated and steadfastly guarded as trade secrets by the respective search engine companies. From our experience and discussion with those involved in the field we would contend with some certainty that robots are guided by a set of criteria that steer them in a systematic way to select certain types of sites and pages, and not select others. However, the blackout on information about search-engine crawl algorithms means we can only try to infer the character of these algorithms from search-engine selection patterns—an inexact exercise.

We have learned something of the nature of spider algorithms from a paper on efficient crawling by Cho, Garcia-Molina, and Page[14], presented at the WWW7 conference (Cho et al., 1998).[15] This paper, which discusses commonly used metrics for determining the "importance" of a Web page by crawling spiders, provides key insights relevant to the main claims of our article. Because of its significance, we discuss it here in some detail. Cho et al (1998) write:

---

[14] Page is one of the designers of Google and the details presented here are the heuristics used by Google (at least the earlier version of these heuristics).

[15] J. Cho, H. Garcia-Molina and L. Page "Efficient crawling through URL ordering" paper presented at the *Seventh International World Wide Web Conference*, 14-18 April 1998, Brisbane, Australia—available at http://www7.scu.edu.au/programme/fullpapers/1919/com1919.htm

Given a webpage P, we can define the importance of the page, I(P), in one of the following ways…:

1. *Similarity to a Driving Query Q. A query Q drives the crawling process,* and *I(P)* is defined to be the textual similarity between *P* and *Q*…
2. *Backlink Count*. The value of I(P) is the number of links to P that appear over the entire web. We use IB(P) to refer to this importance metric. *Intuitively, a page P that is linked to by many pages is more important than one that is seldom referenced.* On the web, IB(P) is useful for ranking query results, giving end-users pages that are more likely to be of general interest. Note that evaluating IB(P) requires counting backlinks over the entire web. A crawler may estimate this value with IB'(P), the number of links to P that have been seen so far.
3. *PageRank.* The IB(P) metric treats all links equally. *Thus, a link from the Yahoo! home page counts the same as a link from some individual's home page. However, since the Yahoo! home page is more important (it has a much higher IB count), it would make sense to value that link more highly. The PageRank backlink metric, IR(P), recursively defines the importance of a page to be the weighted sum of the backlinks to it.* Such a metric has been found to be very useful in ranking results of user queries [Page 1998.2]. We use IR'(P) for the estimated value of IR(P) when we have only a subset of pages available.
4. *Location Metric*. The IL(P) importance of page P is a function of its location, not of its contents. If URL u leads to P, then IL(P) is a function of u. *For example, URLs ending with ".com" may be deemed more useful than URLs with other endings, or URL containing the string "home" may be more of interest than other URLs. Another location metric that is sometimes used considers URLs with fewer slashes more useful than those with more slashes*. All these examples are local metrics since they can be evaluated simply by looking at the URLs." (emphasis added)

The "*Similarity to a Driving Query Q"* metric uses a query term or string (Q)—such as "holiday cottages," for example—as the basic heuristic for crawling. This means that the spider does not need to make a decision about "importance" since it will be directed in its search by the query

string itself. For our discussion, this metric is of minor significance.[16] The real issue emerges when the crawling spider must "decide" importance without the use of a submitted query term. This is where the other metrics play the dominant role. The *Backlink* metric uses the backlink (or inlink) count as its "importance" heuristic. The value of the backlink count is the number of links to the page that appear over the entire Web—for example the number of links over the entire Web that refers to http://www.ibm.com. The assumption here is that "a page that is linked to by many [other] pages is more important than one that is seldom referenced." Obviously, this is a very reasonable heuristic.[17] We know from academic research that it is wise to look at the "canonical" works that is referred to—or cited in academic language—by many other authors. We know also, however, that not all topics necessarily have canons. Furthermore, although in some fields, a small number of citations may make a particular work a canon, in other fields, it takes a vast number of citations to reach canonical status. Thus, the *Backlink* heuristic would tend to crawl and gather the large topics/fields (such as 'shareware computer games') since an even relatively unimportant site in this big field will be seen as more "important"—have relatively more backlinks or inlinks—than an actually important site in a small field (such as "the local community services information"page) which would have relatively less backlinks or inlinks. The essential point is that the large fields determine the measure, or threshold, of "importance"—through sheer volume of backlinks—in ways that would tend to push out the equally important small fields. (We will return to this issue, in our market discussion).

With the *PageRank* metric this problem is exacerbated. Instead of treating all links equally, this heuristic gives prominence to back links from other 'important' pages—pages with high backlink counts. Thus, " since [a link from] the Yahoo! home page is more important (it has a much higher IB [backlink] count), it would make sense to value that link more highly." In the

---

16 We are not claiming that this is a straightforward and uncontroversial metric. The decision about the 'similarity' between the query term and the document is by no means trivial. Decisions how to implement the determination of 'similarity' can indeed be of significant for our discussion. However, we will not pursue this discussion here.

17 In the cases of Excite, Hotbot and Lycos, there is evidence that this is a major consideration for determining indexing appeal—refer http://www.seachenginwatch.com/webmasters/features.html. Exclusion, using this metric is less likely for a search engine like Alta Vista, which goes for massive coverage than for its smaller, more selective, competitors.

analogy to academic papers, a metric like this would imply that a particular paper is even more important if referred to by others whom are already seen as important—by other canons. More simply, you are important if others who are already seen as important indicate that you are important. The problem with the *Backlink* and *PageRank* metrics is that they assume that backlinks are a reliable indication of importance or relevance. In those cases where authors of pages create links to other pages they see as valuable this assumption may be true. There are, however, many organizations that actively cultivate backlinks by inducing Web-page creators to add a link to their page through incentives such as discounts on products, free software utilities, access to exclusive information, and so forth. Obviously not all Web-pages creators have equal access to the resources and incentive to induce others to link to them.

The *Location Metric* uses location information from the URL to determine "next steps" in the crawl. "For example, URLs ending with '.com' may be deemed more useful than URLs with other endings, or URL containing the string 'home' may be more of interest than other URLs." Even though the authors do not indicate what they see as more important, one can assume that these decisions are made when crawl heuristics are set for a particular spider. It may therefore be of great significance "where you are located" as to how important you are seen to be. With the URL as the basis of decision making, many things can aid you in catching the attention of the crawling spider, such as having the right domain name, being located in the root directory, and so forth. From this discussion on crawling metrics we can conclude that pages with many backlinks, especially backlinks from other pages with high backlink counts, which reside in locations seen as "useful" or "important" to the crawling spider, will become targets for harvesting.

Another criterion that seems to guide spiders is breadth or depth of representation. If a spider's algorithm favors breadth (rather than depth), it would visit more sites (or hosts) but index these only partially. In the case of big sites such as America On Line (AOL), Geocities, and so forth, spiders will index them at a rate of approximately 10-15%.[18] If your site is hosted on AOL

---

18 For the search engine operator it is a matter of deciding between breadth and depth. Should they have many sites partially indexed or few sites fully indexed since they know a priori that they can not include everything. Louis Monier, in a response to John Pike—webmaster the Federation of American Scientists site—indicated that Alta Vista indexed 51,570 of the estimated 300,000 pages of the Geocities site. This would amounts to approximately 17% coverage. He thought this to be exceptionally good. Pike indicated that Alta Vista indexed 600 of their 6000

or another big site there is a good chance that it will not be included. Another reason that a site, and so all the pages on that server, may be excluded from search engine databases is that the owner/Webmaster of that server has excluded robots through the robot exclusion standard by means of a "robots.txt" file.[19] This is often done because requests for pages from robots may significantly increase the load on a server and reduce the level of service to all other users. CNN, for example, excludes all robots from its site,[20] as do many sites that offer free Web-page space.[21] It is also important to note that harvesting robots of the search engines we looked at only process HTML files and in particular HTML tags.[22]

Having said all of this it ought to be acknowledged that most spider based search engines do also allow autonomous submissions by Webmasters/designers. Software is available that automatically generates the required electronic formats and facilitates submission to a number of search engines simultaneously. This route has had very mixed results according to the Webmasters we spoke to.

On Being Ranked

---

pages. (refer to this discussion at http://www4.zdnet.com/anchordesk/talkback/talkback_11638.html and http://www4.zdnet.com/anchordesk/talkback/talkback_13066.html as well as to the *New Scientist* paper at http://www.newscientist.com/ keysites/networld/lost.html.

19 For a discussion of this standard, refer to http://info.webcrawler.com/mak/projects/robots/exclusion.html.

20 Another reason for excluding robots from sites such as CNN is that their content is constantly in flux and they do not want sear engines to index (and now cache) old content. Another issue worth noting here is that many search engines now have large cashes to go along with their indexes.

21 Refer to New Scientist paper at http://www.newscientist.com/keysites/networld/lost.html. The 'cost' of a robot visit can be significant for a site. Responsible robots will only request a page every so many seconds. However, the pressure to index has induced what is termed 'rapid fire'. This means that the robot requests in rapid succession that may make the server unavailable to any other user. Although there is a danger that this problem will worsen, there seems to be a generally optimistic view among experts that we will develop technical mechanisms to deal with it, for example, proposals to devise extensions to HTTP, or parallel robots.

22 Although at present some robots are unable to deal with features such as frames, and are better with simple HTML files, there are spiders that have been developed that are now able to handle a variety of formats.

Indexing is but one hurdle to clear for the creators of Web pages who strive for recognition through search engines. Having been successful in the indexing game, their concern shifts to ranking. Many observe that to be noticed by a person doing a search, a Web page needs to be ranked among the top 10-20 listed as hits. Because most search engines display the 10 most relevant hits on the first page of the search results, Web designers jealously covet those ten or twenty top slots. The importance of ranking is regularly discussed by leading authors in the field of web-site promotion:

> There is competition for those top ten seats. There is serious competition. People are trying to take away the top spots every day. They are always trying to fine tune and tweak their HTML code and learn the next little trick. The best players even know dirty ways to "bump off" their competition while protecting their own sites (Anderson & Henderson, 1997).

Although we have not found large-scale empirical studies measuring the effects of ranking on the behavior of seekers, we observe anecdotally that seekers are likely to look down a list and then cease looking when they find a "hit." A study of travel agents using computerized airline reservations systems, which showed an overwhelming likelihood that they would select a flight from the first screenful of search results, is suggestive of what we might expect among Web users at large (Friedman & Nissenbaum, 1996). Indeed if this were not the case it would be difficult to see why Webmasters are going to all the effort to get into the first screen; there is significant evidence that they do indeed take it very seriously. Now it may be that it is not only the first screen but could include the second and third screen as well. Nevertheless, even though we can not say without further research exactly where this line may be (and it may vary with topic, type of searcher, and so forth) we can propose that it does matter whether you are in the first few screens rather than to much lower down the order. One could also argue such a position from an information overload point of view—we shall not pursue it here (Wurman, 1989).

Relevancy ranking is an enormously difficult task. Some researchers working on search technologies argue that relevancy ranking is currently the greater challenge facing search engines and that developments in technical know-how and sheer capacity to find and index sites has not nearly been matched by the technical capacity to resolve relevancy ranking. Besides the engineering challenges, experts must struggle with the challenge of approximating a complex

human value ("relevancy") with a computer algorithm. In other words, according to these experts, while we seem to be mastering the coverage issue, we continue to struggle with the issue of what precisely to extract from the enormous bulk of possibilities for a given search.[23]

Most ranking algorithms of search engines use both the position and the frequency of keywords as a basis for their ranking heuristics (Pringle et al., 1998). Accordingly, a document with high frequency of keywords in the beginning of a document is seen as more relevant (relative to the keyword entered) than one with low frequency lower down in the document. Other ranking schemes, like the heuristic used by Lycos, are based on so-called "in-link" popularity. The popularity score for a particular site is calculated on the total number of other sites that contain links to that site (also refer to backlink value above). High link popularity leads to an improved ranking. As with the crawl metrics discussed above one sees the standard or threshold of "relevance" being set by the big sites at the expense of equally relevant small sites.

The desire and battle for ranking have generated a field of knowledge called "search engine design," which teaches how to design a Web page in order to optimize its ranking and combines these teachings with software to assess its ranking potential. On the one end of the spectrum, practices that make reasonable use of prima facie reasonable heuristics, help designers to optimize their Web pages' expected rankings when they are legitimately relevant to the person searching. On the other end of the spectrum some schemes allow Web designers to manipulate, or trick, the heuristics—schemes such as relevancy (or keyword) spamming,[24] where Web-page designers "trick" the ranking algorithm into ranking their pages higher than they deserve to be ranked by means of keyword stuffing, invisible text, tiny text, and so forth. Such spamming activities doubly punish the innocent. If, for example, you design a Web page with a few graphic images at the beginning, followed somewhere toward the middle with text, you would be severely "punished" by the algorithm both because key terms are positioned relatively low down on the page and also because you would be competing for rank with those less, as it were, scrupulous in their designs.

---

23 Lee Giles disputes this. He still considers indexing to be a huge problem.

24 Also referred to as spamdexing. Refer to http://www.jmls.edu/cyber/index/metatags.html for a reasonable discussion on this issue.

Out of this strange ranking warfare has emerged an impossible situation: Search-engine operators are loath to give out details of their ranking algorithms for fear that spammers will use this knowledge to trick them.[25] Yet, ethical Web-page designers can legitimately defend a need to know how to design for, or indicate relevancy to, the ranking algorithm so that those who search find what is genuinely relevant to their searches.[26]

Beyond the challenge of second-guessing ranking algorithms, there may yet be another, more certain, method of getting results. Some producers of Web sites pursue other ways of elevating their ranking outside of the technical fray: They try to buy them. This subject is an especially sensitive one, and representatives of several major search engines indignantly deny that they sell search positions. Recently, however, in a much-publicized move, Alta Vista and Doublclick have invited advertisers to bid for position in their top slots (Hansell, 1999). Yahoo! sells prominence indirectly by allowing Web owners to pay for express indexing. This allows them to move ahead in the 6-month queue. Another method for buying prominence—less controversial but not unproblematic—allows Web owners to buy keywords for purposes of banner ads. Amazon Books, for example, has a comprehensive arrangement with Yahoo!, and Barnes & Noble with Lycos. If a seeker submits a search to Yahoo! with the term "book" in it, or a term with a name that corresponds to an author's name, or book title in the Amazon database, he would get the Amazon banner (and URL) on his or her search result screen. This is also true for many other companies and products.

The battle for ranking is fought not only between search engines and Web masters/designers but also among organizations wishing for prominence. There is sufficient

---

25 "To stay ahead of the game, the major Search Engines change their methods for determining relevancy rankings every few months. This is usually when they discover that a lot of people have learned the latest technique and are all sneaking into a side door. They also try to fool the tricksters... sometimes they put irrelevant pages at the top of the list just to cause confusion." Patrick Anderson & Michael Henderson, Editor & Publisher, *Hits To Sales* at http://www.hitstosales.com/2search.html.

26 At the WWW7 Conference, researchers in Australia devised an ingenious method for attempting to reverse engineer the relevance ranking algorithms of various commercial search engines causing consternation and some outrage—see G. Pringle, L. Allison and D. L. Dowe, "What is a tall poppy among webpages?" paper presented at the *Seventh International World Wide Web Conference*, 14-18 April 1998, Brisbane, Australia—available at http://www7.scu.edu.au/programme/fullpapers/1872/com1872.htm

evidence to suggest that the fierce competition for both presence and prominence in a listing has led to practices such as one organization retrieving a competitor's Web page, editing it so that it would not do well in the ranking, and resubmitting it as an updated submission, or one organization buying a competitor's name as a keyword and then having the first organization's banner and URL displayed when a search is done on that keyword.[27]

In Table 1 (see Appendix), we summarize the main points of our description, showing some of the ways search engine designers and operators commonly make choices about what to include and exclude in their databases. These choices are embedded in human-interpreted decision criteria, in crawl heuristics and in ranking algorithms.

Implications

We may wonder how all this affects the nature of Web users' experiences. Based on what we have learned so far about the way search engines work, we would predict that information seekers on the Web, whose experiences are mediated through search engines, are most likely to find popular, large, sites whose designers have enough technical savvy to succeed in the ranking game, and especially those sites whose proprietors are able to pay for various means of improving their site's positioning. Seekers are less likely to find less popular, smaller, sites, including those that are not supported by knowledgeable professionals.[28] When a search does yield these sites, they are likely to have lower prominence in rankings.

These predictions are, of course, highly general and will vary considerably according to the keywords or phrases with which a seeker initiates a search, and this in turn, is likely to be

---

27 Lawsuits have been filed by Playboy Enterprises Inc. and Estee Lauder Companies Inc. challenging such arrangements between Excite Inc. and other companies that have "bought" their respective names for purposes of banner ads. See Kaplan, C. "Lawsuits Challenge Search Engines' Practice of 'Selling' Trademarks" The New York Times on the web, February 12, 1999. http://www.nytimes.com/library/tech/99/02/cyber/cyberlaw/12law.html

28 "If you want the traffic and the exposure, *you are going to pay for the education or you are going to pay for the service.* There is no other way to do it. It is not easy. It is not magic. It takes time, effort, and knowledge. Then it takes continual monitoring to keep the position you worked so hard to get in the first place. Please do not misunderstand—the competition is fierce and severe for those top spots, which is why the Search Engines can charge so much money to sell keyword banners." Patrick Anderson & Michael Henderson, Editor & Publisher, Hits To Sales at http://www.hitstosales.com/2search.html. (emphasis added)

affected by the seeker's competence with search engines. The nature of experiences of information seekers will also vary according to the search engines they choose. Some users may actively seek one search engine over others but some will simply, and perhaps unknowingly, use a default engine provided by institutions or Internet service providers (ISPs).[29] We are unlikely to find much relief from these robust regularities from meta search engines, like Metacrawler, Ask Jeeves, and Debriefing, because they base their results on existing search engines and normally accomplish their task by recognizing only higher order search keys than first-order engines.[30] We note further that not only are most users unaware of these particular biases, they seem also to be unaware that they are unaware.

SHOULD WE LET THE MARKET DECIDE?

Readers may find little to trouble them in this description of search engine proclivities. What we have before us is an evolving marketplace in search engines: We ought to let producers of search engines do what they will and let users decide freely which they like best. Search engines whose offerings are skewed either because their selections are not comprehensive, or because they prioritize listings according to highest bid, will suffer in the marketplace. And even if they do not, the collective preferences of participants should not be second-guessed. As the representatives of Yahoo! we cited earlier have argued, users' reactions must remain the benchmark of quality: Dissatisfied seekers will defect from an inadequate search engine to others that do a better job of indexing and prioritizing. Thus will the best search engines flourish; the poor ones will fade away from lack of use. McChesney (1997b) describes a comparable faith in the market mechanism as it applied to traditional broadcast media: "In the United States, the notion that commercial

---

29 Some large sites (universities, for example) allow users to submit keywords, which the site, in turn, submits to a particular default search engine (frequently, Yahoo!). If users select 'search' on the Netscape toolbar it takes them to the Netscape webpages where they have a list of search engines. In this case Excite is the default search engine. There is clearly considerable advantage to being chosen as the default search engine on the Netscape, or other equivalent, webpage.

30 This is because, as Giles and Lawrence remarked in verbal consultation, there is a fair degree of convergence in the results yielded by various search engine algorithms and decision criteria.

broadcasting is the superior system because it embodies market principles is closely attached to the notion that the market is the only 'democratic' regulatory mechanism, and that this democratic market is the essence of Americanism, patriotism, and all that is good and true in the world." Both McChesney (1999) and Schiller (1995), however, have criticized the idea that a media market best represents democratic ideals. In the case of search engines, we are, likewise, not optimistic about the promise of development shaped only by a marketplace.

As anyone who has used search engines knows, the dominant search engines do not charge seekers for the search service. Rather, the arrangement resembles that of commercial television where advertisers pay television stations for the promise of viewers. Similarly, search engines attract paid advertisements based on the promise of search usage. High usage, presumably, garners advertisers and high charges. To succeed, therefore, search engines must establish a reputation for satisfying seekers' desires and needs; this way they will attract seekers in the first place, and then will keep them coming back[31]. As a way of simplifying the discussion, however, we will refer to the marketplace as a marketplace in search engines with users, or seekers, as the buyers. This strategy does not, as far as we have been able to tell, alter the substantive outcomes of the particular issues we have chosen to highlight.

We do not dispute the basic fact of the matter, namely, that a marketplace for search engines (and seekers, if you will) is possible. It is also possible that such a market, reflecting discrepant degrees of satisfaction by seekers, will result in some search engines flourishing and others failing. Our dissatisfaction with this forecast is not that it cannot come true, but what it would mean, from the perspective of social values and the social investment in the Internet, if it did. Why, the critic might ask, on what grounds, would we presume to override the wishes of users so cleanly reflected in their market choices? Our reply to this challenge, which we try to keep as free from sentimental prejudices as possible, cites two main sources of concern. One is that the conditions needed for a marketplace to function in a "democratic" and efficient way are simply not met in the case of search engines. The other is our judgement that Web search mechanisms are too important to be shaped by the marketplace alone. We discuss each in turn, the first one only briefly.

A virtue frequently claimed by defenders of the market mechanism is that participants are free to express their preferences through the choices they make among alternatives. Through their choices, incompetent inefficient suppliers are eliminated in favor of competent efficient suppliers. As many critics have pointed out, however, this holds only for markets where those who supply goods or services have an equal possibility to enter the market and communicate with potential customers, and those who demand goods and services are fully informed and will act in a rational manner. Such an ideal market simply does not exist, and this is especially so in the case of search engines.

If we focus on the demand side first, we see that most users of the Web lack critical information about alternatives. Only a small fraction of users understand how search engines work and by what means they yield their results. It is misleading to suggest that these users are meaningfully expressing preferences or exercising free choice when they select from the alternatives. Though we lack systematic empirical evidence, the anecdotal results of asking people why they use or prefer one search engine to others is some version of, "It finds what I'm looking for," and a shrug. Now if one is searching for a specific product or service it may be possible to know in advance how to determine that one has indeed found what one was looking for. When searching for information, however, it is difficult (if not impossible) to make such a conclusive assessment, since the locating of information also serves to inform one about that which one is looking for. This is an old information retrieval problem—often expressed as "how do you know what you do not know until you know it"—with which information science scholars have been battling for many years. It seems unlikely that this would be different for search engines. In fact, the partiality of any search attempt (even if we assume a competent searcher) will magnify this problem in the search-engine context. Not only this, we would also claim that users tend to be ignorant about the inherent partiality present in any search engine search results (as explained earlier, in the technical overview above). They tend to treat search-engine results the way they treat library catalogue search results. Given the vastness of the Web, the close guarding of algorithms, and the abstruseness of the technology to most users, it should come as no surprise that seekers are unfamiliar, even unaware, of the systematic mechanisms that drive search engines.

---

[31] One should also mention that search engines also market themselves aggressively. They also establish agreements with other service providers to become defaults on their pages. Refer to

Such awareness, we believe, would make a difference. Although here too we came across no systematic empirical findings, we note that in spheres outside of the electronic media, people draw clear and definitive distinctions between information and recommendations coming from disinterested, as compared with interested, sources, impartial advice as compared with advertisement.[32] And anecdotal experience bears this out as when customers learned that Amazon Books, for example, had been representing as "friendly" recommendations what were in reality paid advertisements. Customers responded with great ire and Amazon hastily retreated.

The problem is equally complex on the supply side of the supposed market. We have indicated the complex hurdles that need to be cleared to get listed and ranked appropriately. They all indicate that there simply is no level playing field by any stretch of the imagination. It seems clear that the "market will decide" view (problematic in most cases) is extremely problematic in this context. It is also doubtful weather this can be resolved to the point that the market argument will become valid.

The question of whether a marketplace in search engines approximates sufficiently a competitive free market is, perhaps, subordinate to the question whether we ought to leave the shaping of search mechanisms to the marketplace in the first place. We think this would be a bad idea.

Developments in Web searching are shaped by two distinct forces. One is the collective preferences of seekers. On the current, commercial model, search engines wishing to achieve greatest popularity would tend to cater to majority interests. While markets undoubtedly would force a degree of comprehensiveness and objectivity in listings, there is unlikely to be much market incentive to list sites of interest to small groups of individuals, such as individuals interested in rare animals or objects, individuals working in narrow and specialized fields, or, for that matter, individuals of lesser economic power, and so forth. But popularity with seekers is not the only force at play. The other, is the force exerted by entities wishing to be found. Here, there is enormous inequality. Some enter the market already wielding vastly greater prowess and economic power than others. The rich and powerful clearly can influence the tendencies of search engines; their dollars can (and in a restricted way do already) play a decisive a role in what gets

---

footnote 33 above.

[32] As noted by one of the reviewers this is equally less true outside the electronic media.

"found." For example, of the top 100 sites—based on traffic—just 6 are not ".com" commercial sites[33]. If we exclude universities, NASA and the U.S. government this drops to two. One could reasonably argue that the United Nations site ought to generate at least enough traffic to be in the list if we consider that Amazon is in position number 10 and *USA Today* in position 35.[35] The cost to a search engine of losing a small number of searching customers may be outweighed by the benefits of pandering to "the masses" and to entities paying fees for the various forms of enhanced visibility. We can expect, therefore, that at least some drift will be caused by those wishing to be found which, in turn, would further narrow the field of what is available to seekers of information, association, support, and services.[34]

It may be useful to think of the Web as a market of markets instead of as just one market. When we seek, we are not interested in "information" in general; rather we are interested in specific information related to our specific interests and needs. Seekers might be in the market for information about, for example, "packaged tour holidays" or "computer hardware suppliers." For these markets, where we expect the demand for information to be great, we would expect the competition for recognition to be great as well. Companies would pay high prices for the keyword banners that will ensure them the "top spot" and a search will generate many "hits" for the seekers. In contrast, there are other significantly smaller markets—for information about a rare medical condition, or about services of a local government authority or community.

In this market of markets, there is likely to be little incentive to ensure inclusion for these small markets, and only a small cost (in loss of participation) for their exclusion. Although we do not have empirical evidence, we would expect the law of Pareto to apply (see Sen, 1985). We could imagine that a high percentage of search requests (say 80%, for argument's sake) are directed to a small percentage (say 20%) of the big markets, which would be abundantly represented in search results.[35] Only a small percentage of the search requests (say 20%) might be addressed to the large percentage (say 80%) of the smaller markets, which would be

---

[33] Refer to http://www.100hot.com for latest list.

[34] And engines that use link popularity for priority listing will be even more prone to reifying a mode of conservatism on the web.

[35] This guess is not far off from reality as searches for sex-related key terms are by far the most frequent—constituting perhaps as high a percentage as 80% of overall searches.

underrepresented. This scenario would explain the limited incentive for inclusion and relatively low cost of exclusion. We find this result problematic.

A market enthusiast does not find this result problematic. This is exactly what the market is supposed to do; the range and nature of choices are supposed to ebb and flow in response to the ebb and flow of wants and needs of market participants—from varieties of salad dressings to makes of automobiles. Nevertheless, we resist this conclusion not because we are suspicious of markets in general—for cars and salad dressings they are fine—but because maintaining the variety of options on the Web is of special importance. We resist the conclusion because we think that the value of comprehensive, thorough and wide-ranging access to the Web lies within the category of goods that Elizabeth Anderson describes in her book, *Values in Ethics and Economics* as goods that should not be left entirely (if at all) to the marketplace (Anderson, 1993).

Anderson constructs an elaborate argument defending the claim that there are ethical limitations on the scope of market norms for a range of goods (and services). Abstracting principles from cases that are likely to be uncontroversial in this regard, for example, friendship, persons, and political goods (like the vote), she then argues that these principles apply to goods that are likely to be more controversial in this regard such as public spaces, artistic endeavor, addictive drugs, and reproductive capacities. For some goods, such as cars, bottled salad dressings, etc., "unexamined wants," expressed through the marketplace, are a perfectly acceptable basis for distribution. For others, including those that Anderson identifies, market norms do not properly express the valuations of a liberal democratic society like ours, which is committed to "freedom, autonomy and welfare" (Anderson, 1993, p.141). Although it is not essential to our position that we uncritically accept the whole of Anderson's analysis, we accept at least this: that there are certain goods—ones that Anderson calls "political goods," including among them schools and public places—that must be distributed not in accordance with market norms but "in accordance with public principles" (Anderson, 1993, p.159).

Sustaining the 80% of small markets that would be neglected by search engines shaped by market forces qualifies as a task worthy of public attention. Sustaining a full range of options here is not the same as sustaining a full range of options in bottled salad dressings, or cars, because the former enriches the democratic arena, may serve fundamental interests of many of the neediest

members of our society, and more (on which we elaborate in the next section). We make political decisions to "save" certain goods that might fall by the wayside in a purely market-driven society. In this way, we recognize and save national treasures, historic homes, public parks, schools, and so forth. In this spirit, we commit to serving groups of people, like the disabled, even though (and because) we know that a market mechanism would not cater to their needs. (We make special accommodation for non-profit efforts through tax-exemption without consideration for popularity). We see an equivalent need in the case of search engines.

In order to make the case convincing, however, we need to introduce into the picture a substantive claim because our argument against leaving search engines fully to the mercy of the marketplace is not based on formal grounds—or at least, we do not see them. We base our case against the "leave it to the market challenge" on the particular function that we see search engines serving and on the substantive vision of the Web that we think search engines (and search and retrieval mechanisms more generally) ought to sustain. We do not argue unconditionally that the trajectory of search engine development is wrong or politically dangerous in itself, but rather that it undermines a particular, normative vision of the Web in society. Those who do not share in this vision are unlikely to be convinced that search engines are different (in kind) from salad dressings and automobiles. The case that search engines are a special, political good presumes that the Web, too, is a special good.

THE FUTURE OF THE WEB[36] AS A PUBLIC GOOD

The thesis we here elaborate is that search engines, functioning in the manner outlined earlier, raise political concerns not simply because of the way they function, but because the way they function seems at odds with the compelling ideology of the Web as a public good. This ideology portrays the fundamental nature and ethos of the Web as a public good of a particular kind, a rich array of commercial activity, political activity, artistic activity, associations of all kinds, communications of all kinds, and a virtually endless supply of information. In this regard the Web

36 Our discussion of the Web would probably be more accurately addressed to the Internet as a whole. We think that the more inclusive discussion would only strengthen our conclusions but would probably introduce unnecessary complexity.

was, and is still seen by many as a democratic medium that can circumvent the hegemony of the traditional media market, even government control.

Over the course of a decade, or so, computerized networks—the Internet and now the Web—have been envisioned as a great public good. Those who have held and promoted this vision over the course of, perhaps, a decade, have based their claims on a combination of what we had already achieved and what the future promises. For example, with only a fraction of the population in the U.S. linked to the Internet, Al Gore (1995) promoted the vision of a Global Internet Infrastructure. This conception of the great public good—part-reality, part wishful thinking—has gripped people from a variety of sectors, including scholars, engineers and scientists, entrepreneurs and politicians. Each has highlighted a particular dimension of the Web's promise, some focusing on information, some on communication, some on commerce, and so on. Although we cannot, here, enumerate all possible public benefits, we highlight a few.

A theme that is woven throughout most versions of the promise is that the Web contributes to the public good by serving as a special kind of public space. The Web earns its characterization as public in many of the same ways as other spaces earn theirs, and contributes to the public good for many of the same reasons. One feature that pushes something into the realm we call public is that it is not privately owned. The Web does seem to be public in this sense: its hardware and software infrastructure is not wholly owned by any person or institution, or, for that matter, by any single nation. Arguably, it does not even come under the territorial jurisdiction of any existing sovereign state.[37] There is no central or located clearinghouse that specifies or vets content, or regulates overall who has the right of access. All those who accept the technical protocols, conform to technical standards (HTML, for example) and are able to connect to it may "enter" the Web. They may access others on the Web, and, unless they take special precautions, they may be accessed. When I post my Web pages, I may make them available to any of the millions potential browsers, even if, like a street vendor, I decide to charge a fee for entry to my page. The collaborative nature of much of the activity on the Web leads to a sense of the Web not simply as unowned but as collectively owned.

---

37 David R. Johnson and David Post, "Law and Borders—The Rise of Law in Cyberspace," *Stanford Law Review*, Volume 48, No. 5, May 1996, 1367-1402. This article puts forward an extreme version of this view. We will not engage further in the debate.

The Web fulfills some of the functions of other traditional public spaces—museums, parks, beaches, and schools. It serves as a medium for artistic expression, a space for recreation, a place for storing and exhibiting items of historical and cultural importance, and it can educate. Beyond these functions, the one that has earned it greatest approbation as both a public space and a political good is its capacity as a medium for intensive communication among and between individuals and groups in just about all the permutations that one can imagine, namely, one-to-one, one-to-many, etc. It is the Hyde Park Corner of the electronic age, the public square where people may gather as a whole, or associate in smaller groups. They may talk and listen, they may plan and organize. They air viewpoints and deliberate over matters of public importance. Such spaces, where content is regulated by only a few fundamental rules, embody the ideals of the liberal democratic society.

The idea of the Web as a public space and a forum for political deliberation has fueled discussions on teledemocracy for some time (Abramson et al., 1988; Arterton, 1987). The notion of the public sphere as a forum in which communicatively rational dialogue can take place unsullied by ideology has had one of its strongest proponents in Habermas (1989). Although there is no universal agreement among scholars on the extent of effects the Web may have in the political sphere, several contributors to the debate have cited cases in which the Web appears to have had a decisive impact on the outcome. Douglas Kellner (1997) gives some examples: Zapatistas in their struggle against the Mexican government, the Tiananmen Square democracy movement, environmental activists who exposed McDonald's through the McLibel campaign, and the Clean Clothes Campaign supporting attempts of Filipino garment workers to expose exploitative working conditions.[38]

We have not yet mentioned perhaps the dominant reason for conceiving of the Web as a public good, namely, its function as a conveyor of information. As a public means of access to vast amounts of information, the Web promises widespread benefits. In this so-called "information age," being among the information-rich is considered to be so important that some, like the philosopher Jeroen van den Hoven (1994, 1998) have argued that it makes sense to construe access to information as one of the Rawlsian "primary goods," compelling any just society to

38 Refer to his paper "Intellectuals, the New Public Spheres, and Techno-Politics" at http://www.gseis.ucla.edu/ courses/ed253a/newDK/intell.htm

guarantee to a basic, or reasonable, degree of it to all citizens. Growing use of the Web as a repository for all manner of information (e.g., government documents, consumer goods, scientific and artistic works, local public announcements, etc.) lends increasing weight to this prescription. The Web, according to the vision, is not intended as a vehicle for further expanding the gap between haves and have-nots, but for narrowing it (see, e.g., Civille, 1996; Hoffman & Novak, 1998).

The view of the Internet as a public good, as a globally inclusive, popular medium fueled much of the initial social and economic investment in the medium and its supporting technology, convincing progressive politicians (or those who wished to appear progressive) to support it with investment and political backing.[39] The vision has also motivated idealistic computer scientists and engineers to volunteer energy and expertise toward developing and promulgating the hardware and software, from the likes of Jonathan Postel, one the early builders of the Internet, who worked to keep its standards open and free,[40] to professionals and researchers volunteering in efforts to wire schools and help build infrastructure in poorer nations. These inclusive values were very much in the minds of creators of the Web like Tim Berners-Lee:

> The universality of the Web includes the fact that the information space can represent anything from one's personal private jottings to a polished global publication. We as people can, with or without the Web, interact on all scales. By being involved on every level, we ourselves form the ties which weave the levels together into a sort of consistency, balancing the homogeneity and the heterogeneity, the harmony and the diversity. We can be involved on a personal, family, town, corporate, state, national, union, and international levels. Culture exists at all levels, and we should give it a weighted balanced respect at each level.[41]

---

39 Popular news media reflect the hold of this vision of the web. In an article in the New York Times about the Gates Learning Foundation's recent donation for public-access computers to libraries, the gift is discussed in terms of bridging economic inequality and overcoming technical illiteracy. Librarians are quoted as enthusiastically reporting that the computers are uses, "to type (their) resumes, hunt for jobs, do schoolwork, research Beanie Babies, look up medical information, investigate their family roots, send E-mail and visit wrestling sites on the web." Katie Hafner, *The New York Times*, February 21, 1999.

40 "A Net Builder Who Loved Invention, Not Profit," *The New York Times*, October 22, 1998.

41 Refer to http://www.w3.org/1998/02/Potential.html

While the promise of the Web as a public space and a public good continues to galvanize general, political, and commercial support, many observers and scholars have cautioned that the goods are not guaranteed. The benefits of the vast electronic landscape, the billions of gigabytes of information, and the participation of millions of people around the world, depend on a number of contingencies. Issuing one such caution, Lewis Branscomb (1996) calls for political effort to protect public interests against encroaching commercial interests. He worries about the enormous amount of money "invested in the new business combinations to exploit this consumer information market; the dollars completely swamp the modest investments being made in bringing public services to citizens and public institutions," (p.27), urging federal, state and local government to "develop and realize the many non-profit public service applications necessary for the realization of the 'promise of NII'" (p.31).

Gary Chapman and Marc Rotenberg, writing in 1993 on behalf of the organization Computer Professionals for Social Responsibility, listed a number of problems that would need to be solved before the National Information Infrastructure would be capable of serving the public interest. Of particular relevance to us here is Chapman and Rotenberg's reference to Marvin Sirbu's (1992) call for "Development of standardized methods for information finding: White Pages directories, yellow Pages, information indexes." Without an effective means of finding what you need, the benefits of an information and communication infrastructure like the Web are significantly diminished. We can conjure up analogies: a library containing all the printed books and papers in the world without covers and without a catalogue; a global telephone network without a directory; a magnificent encyclopaedia, haphazardly organized and lacking a table of contents.

Search engines are not the only answer to this need but they still are the most prominent, the one to which most users turn when they want to explore new territory on the web. The power, therefore, that search engines wield in their capacity to highlight and emphasize certain Web sites, while making others, essentially, disappear, is considerable. If search engines systematically highlight Web sites with popular appeal and mainstream commercial purpose, as well as Web sites backed by entrenched economic powers, they amplify these presences on the Web at the expense of others. Many of the neglected venues and sources of information,

suffering from lack of traffic, perhaps actually disappear, further narrowing the options to Web participants.

If trends in the design and function of search engines leads to a narrowing of options on the Web—an actual narrowing or a narrowing in what can be located, the Web as a public good of the particular kind that many envisioned is undermined. The ideal Web serves all people, not just some, not merely those in the mainstream. It is precisely the inclusivity and breadth that energized many to think that this technology would mean not just "business as usual" in the electronic realm, not merely a new tool for entrenched views and powers. The ideal Web would extend the possibilities for association, would facilitate access to obscure sources of information, would give voice to many of the typically unheard, and would preserve intensive and broadly inclusive interactivity.

In considering the effects of a biased indexing and retrieval system, our attention first was drawn to the seekers. It is from the perspective of seekers that we noted the systematic narrowing of Web offerings: there would be fewer opportunities to locate various types of information, individuals and organizations, a narrowing of the full range of deliberative as well as recreational capabilities. If access to the Web is understood as access by seekers to all of these resources, then the outcome of biased search engines amounts to a shrinking of access to the web. This perspective, however, does not represent all that is at stake. At stake is access to the Web in the shape of those, in addition, who would like to be found, to be seen and heard. Marc Raboy describes this dimension of the new medium:

> The notion of "access" has traditionally meant different things in broadcasting and in telecommunications. In the broadcasting model, emphasis is placed on the active receiver, on free choice, and access refers to the entire range of products on offer. In the telecommunications model, emphasis is on the sender, on the capacity to get one's messages out, and access refers to the means of communication. In the new media environment, public policy will need to promote a new hybrid model of communication, which combines the social and cultural objectives of both broadcasting and telecommunications, and provides new mechanisms—drawn from both traditional models—aimed at maximizing equitable access to services and the means of communication for both senders and receivers (Raboy, 1998, p. 224).

The public good of the Web lies not merely in its functioning as a repository for seekers to find things, but as a forum for those with something (goods, services, viewpoints, political activism, etc.) to offer. The cost of a biased search-and-retrieval mechanism may even be greater for Web site owners wishing to be found—the senders. Consider an example of just one type of case, someone seeking information about, say, vacation rentals in the Fiji Islands. Because one rental is all he needs he is likely to look down a list of options and stop looking when he or she finds it. There is no loss to the seeker even if it turns out that lower down on the list there are many other candidates meeting his criteria. He has found what he or she needs. Those who are not found (because their lower ranking deprives them of attention or recognition) are offering, arguably, just as much value to the seeker. Our loss, in this case is twofold: One is that if continuing invisibility causes options to atrophy, the field of opportunity is thinned; the other is that many of those reaching out for attention or connection are not being served by web. If search mechanisms systematically narrow the scope of what seekers may find and what sites may be found, they will diminish the overall value of the Web as a public forum, as well as a broadly inclusive source of information.

Many have observed that to realize the vision of the Web as a democratizing technology, or, more generally, as a public good, we must take the question of access seriously. We agree with this sentiment but would wish to expand what it covers. Access is not merely a computer and a network hookup, as some have argued, nor in addition, the skills and know-how that enables effective use. Access implies a comprehensive mechanism for finding and being found. It is in this context that we raise the issue of the politics of search engines—a politics that at present seems to push the Web into a drift that does not resonate with one of the historically driving ideologies.[42] We also believe we have shown why a rally to the market will not save the day, will not assure our grand purpose. The question of how to achieve it is far harder.

SOME CONCLUSIONS AND IMPLICATIONS

42 Larry Lessig has argued that there has been an unacknowledged but significant shift in this ethos. See "The Law of the Horse: What Cyberlaw Might Teach," *Harvard Law Review* (1999) Forthcoming.

We have claimed that search-engine design is not only a technical matter but also a political one. Search engines are important because they provide essential access to the Web both to those with something to say and offer as well as to those wishing to hear and find. Our concern is with the evident tendency of many of the leading search engines to give prominence to popular, wealthy, and powerful sites at the expense of others. This they do through the technical mechanisms of crawling, indexing, and ranking algorithms, as well as through human-mediated trading of prominence for a fee. As long as this tendency continues, we expect these political effects will become more acute as the Web expands.

We regret this tendency not because it goes against our personal norms of fair play but because it undermines a substantive ideal—the substantive vision of the Web as an inclusive democratic space. This ideal Web is not merely a new communications infrastructure, offering greater bandwidth, speed, massive connectivity, and more, but a platform for social justice. It promises access to the kind of information that aids upward social mobility; it helps people make better decisions about politics, health, education, and more. The ideal Web also facilitates associations and communication that could empower and give voice to those who, traditionally, have been weaker and ignored. A drift toward popular, commercially successful institutions, through the partial view offered by search engines, seriously threatens these prospects. Scrutiny and discussion are important responses to these issues but policy and action are also needed—to fill that half-empty portion of the cup. We offer preliminary suggestions, calling for a combination of regulation through public policy, as well as value-conscious design innovation.

The tenor of our suggestions is toward enhancement. We do not see that regulating and restricting development of commercial search engines is likely to produce ends that we would value—as it were, siphoning off from the half-full portion. This course of action is likely to be neither practically appealing nor wise, and might smack of cultural elitism or paternalism. Amartya Sen (1987, p.9), commenting on existing schools of thought within the field of economics, wrote: "It is not my purpose to write off what has been or is being achieved, but definitely to demand more."[43] We take a similar stance in response to our study of Web search engines.

Policy

As a first step we would demand full and truthful disclosure of the underlying rules (or algorithms) governing indexing, searching, and prioritizing, stated in a way that is meaningful to the majority of Web users. Obviously, this might help spammers. However, we would argue that the impact of these unethical practices would be severely dampened if both seekers and those wishing to be found are aware of the particular biases inherent in any given search engine. We believe informing users, on the whole, will be better than the status quo, in spite of the difficulties. Those who favor a market mechanism would perhaps be pleased to note that disclosure would move us closer to fulfilling the criteria of an ideal competitive market in search engines. Disclosure is a step in the right direction because it would lead to a clearer grasp of what is at stake in selecting among the various search engines, which in turn should help seekers make more informed decisions about which search engines to use and trust. But disclosure, by itself, may not sustain and enhance Web offerings in the way we would like—that is, by retaining transparency for those less popular sites to promote inclusiveness.

The marketplace alone, as we have argued, is not adequate. As a policy step, we might, for example, consider public support for developing more egalitarian and inclusive search mechanisms, and for research into search and meta-search technologies that would increase transparency and access. Evidently, if we leave the task of charting the Web in the hands of commercial interests alone, we will merely mirror existing asymmetries of power in the very structure of the Web (McChesney, 1999). Although these and other policies could promise a fairer representation of Web offerings, a second key lies in the technology itself.

Values in Design

Philosophers of technology have recognized the intricate connection between technology and values—social, political, and moral values.[44] These ideas – that technological systems may embed or embody values—resonate in social and political commentary on information technology written by engineers as well as by philosophers and experts in cyberlaw (see, e.g., Friedman, 1997; Lessig, 1999; Nissenbaum, 1998). Translating these ideas into practice implies that we can build

---

43 Sen, A. *On Ethics and Economics*, Oxford: Blackwell, 1987. p. 9

44 See, for example, Winner, L. 1980. "Do Artifacts Have Politics?" Daedelus 109: 121-136

better systems; that is to say, systems that better reflect important social values, if we build them with an explicit commitment to values. The commitment we hope to inspire with this paper, among the designers and builders of search engine technology, is to the value of fairness as well as to the suite of values represented in the ideology of the Web as a public good.

Two technical approaches that appear to be attracting interest are not without drawbacks. One would increase segmentation and diversification. Search engines would become associated with particular segments of society—borders perhaps drawn according to traditional categories (sport, entertainment, art, and so forth). A problem with segmentation overall, however, is that it could fragment the very inclusiveness and universality of the Web that we value. The Web may eventually merely mirror the institutions of society with its baggage of asymmetrical power structures, privilege, and so forth.

The other approach is to develop individualized robots that go out and search for pages based on individual criteria, building individualized databases according to individual needs.[45] There is, however, a significant "cost" in automatic harvesting via robots that even the existing population of robots imposes on system resources; this has already caused concern (Kostner, 1995).

There is much interesting work underway on the technology of search engines that could, in principle, help: for example, improving the way individual pages indicate relevance (also referred to as metadata) (see Marchiori, 1998), refining overall search engine technology,[46] and improving Web resource presentation and visualization (see Hearst, 1997) and meta-search technology (see Lawrence & Giles, 1998). Although improvements like these might, accidentally, promote values, they hold greatest promise as remedies to the current politics of search engines if they are explicitly guided by values. We urge engineers and scientists who adhere to the ideology of the Web, to its values of inclusivity, fairness, and scope of representation, and so forth, to pursue improvements in indexing, searching, accessing, and ranking with these values firmly in

---

45 Individualized spiders such as *NetAttaché* are already available for as little as $50.00. Refer to http://www.tympani.com/store/NAProTools.html

46 Some cite Google as an example. This is a particularly interesting case as Google started out as a search engine developed within an educational setting and moved into the for-profit sector. We think it would be very worthwhile to trace changes in the technology that might result from this move.

their sights. It is well to keep in mind that the struggle to chart the Web and capture the attention of the information seekers is not merely a technical one, it is also a political one.

Acknowledgements

We are indebted to many colleagues for commenting on and questioning earlier versions of this paper: Audiences at the conference, Computer Ethics: A Philosophical Enquiry, London; members of the seminars at the Kennedy School of Government, Harvard University, and the Center for Arts and Cultural Policy Studies, Princeton University; Steven Tepper, Eszter Hargittai, Phil Agre; Rob Kling and reviewers for The Information Society. We are grateful to Lee Giles, Brian LaMacchia, Andrea LaPaugh (and members of her graduate seminar), and Andrew Tomkins for technical guidance, and to our able research assistants Michael Cohen and Sayumi Takahashi. Nissenbaum acknowledges the invaluable support of the National Science Foundation through the grant SBR-9806234.

Table: Summary of Criteria for Indexing and Ranking

(A) Search Engine Perspective	*Reason for exclusion*
Indexing	
Director- type search engines	(1) The human editor does not include your submission based on criteria not generally known and apparently inconsistently applied.
Automatic-harvesting type search engines	(1) Site not visited because of robot exclusion standard set by the Webmaster.
	(2) Site not in the crawl path of the robot (not sufficiently rich in backlinks).
	(3) Part of a large (often free) site that is only partially indexed.
	(4) Documents don't conform to HTML standard (pdf, gif, etc.).
Ranking (in top 10 *when relevant*)	
	(1) Did not buy the keyword or top spot.
	(2) Not high in inlink popularity (from and to site).
	(3) Relevant keywords not in meta tag or title.
	(4) Keyword spammers have pushed you down.
	(5) Important part of your title are stop words.
	(6) Your pages have been altered (dumped off) through unethical practices by you competitors.
(B) Seeker Perspective: Finding appropriate content	(1) Using only one search engine (sometimes a default that user is unaware of).
	(2) Inappropriate use of search criteria.

REFERENCES

Abramson, Jeffrey B., Arterton, F. C., and Orren, G. R.. 1988. *The Electronic Commonwealth: The Impact of New Media Technologies on Democratic Politics.* New York: Basic Books.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge and London: Harvard University Press.

Anderson, Patrick and Henderson, Michael. 1997. *Hits To Sales.* <http://www.hitstosales.com/2search.html>

Arterton, F. Christopher. 1987. *Teledemocracy: Can Technology Protect Democracy.* Newbury Park, CA: Sage.

Brake, David. 1997. Lost in cyberspace. *New Scientist* June 28. <http://www.newscientist.com/keysites/networld/lost.html>

Branscomb, Lewis. 1996. Balancing the commercial and public-interest visions. In *Public Access to the Internet*, eds. Brian Kahin and James Keller. Cambridge, MA: MIT Press.

Chapman, Gary and Rotenberg, Marc. 1993. The national information infrastructure: A public interest opportunity. *CPSR Newsletter 11(2): 1-23.*

Cho, J., Garcia-Molina H., and Page, L. 1998. Efficient crawling through URL ordering. *Seventh International World Wide Web Conference*, Brisbane, Australia, 14-18 April.

Civille, Richard. 1996. The Internet and the poor. In *Public Access to the Internet*, eds. Brian Kahin and James Keller. Cambridge, MA: MIT Press.

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

Golding, Peter. 1994. The communications paradox: Inequality at the national and international levels. *Media Development* 4: 7-9.

Gore, Al. 1995. Global Information Infrastructure. In *Computers, Ethics and Social Values*, Eds. D. Johnson and H. Nissenbaum. Englewood Cliffs, NJ: Prentice Hall.

Habermas, Jurgen. 1989. *The structural transformation of the public sphere*, Trans. T. Burger and F. Lawrence. Cambridge, MA: Harvard University Press.

Hansell, S. 1999. Altavista invites advertisers to pay for top ranking. *New York Times* 15 April.

Hearst, Marti. 1997. Interfaces for searching the Web. *Scientific American* March. <http://www.sciam.com/0397issue/039/hearst.html>

Hoffman, Donna L., and Novak, Thomas P. 1998. Bridging the racial divide on the Internet. *Science* 280: 390-391.

Johnson, David R., and Post, David. 1996. Law and borders- The rise of law in cyberspace. *Stanford Law Review* 48(5).

Kaplan, C. 1999. Lawsuits challenge search engines' practice of "selling" trademarks. *The New York Times* 12 February. <http://www.nytimes.com/library/tech/99/02/cyber/cyberlaw/12law.html>

Kellner, Douglas. 1997. Intellectuals, the new public spheres, and techno-politics. <http://www.gseis.ucla.edu/courses/ed253a/newDK/intell.htm>.

Kostner, Martijn. Robots in the web: threat or treat. <http://info.webcrawler.com>

Lawrence, S., and Giles, C.L. 1998. Inquirus, the NECI meta search engine. *Seventh International World Wide Web Conference*, Brisbane, Australia, 14-18 April. <http://www7.scu.edu.au/programme/fullpapers/1906/com1906.htm>

Lawrence, S., and Giles, C.L. 1999. Accessibility and Distribution of Information on the Web. *Nature* 400: 107-109.

Lessig, Lawrence 1999. *Code and Other Laws of Cyberspace*. New York: Basic Books.

McChesney, Robert W. 1993. *Telecommunications, Mass Media and Democracy*. Oxford: Oxford University Press.

McChesney, Robert W. 1997a. The Mythology of Commercial Media and the Contemporary Crisis of Public Broadcasting. *Spry Memorial Lecture 1997*, Montreal & Vancouver, 2 & 4 December.

McChesney, Robert W. 1999. Making Media Democratic. *Boston Review: New Democracy Forum.* <http:polisci.mit.edu/BostonReview/BR23.3/mcchesney.html>

McChesney, Robert W. 1997b. *Corporate Media and the Threat to Democracy.* New York: Seven Stories Press.

McChesney, Robert W. and Herman, Edward S. 1997. *The Global Media: The New Missionaries of Corporate Capitalism.* London: Cassell.

Marchiori, M. 1998. The limits of Web metadata, and beyond. *Seventh International World Wide Web Conference*, Brisbane, Australia, 14-18 April. <htpp://www7.scu.edu.au/programme/fullpapers/1896/com1896.htm>

Miller, R. C., and Bharat, K. 1998. SPHINX: a framework for creating personal, site-specific Web crawlers. *Seventh International World Wide Web Conference*, Brisbane, Australia, 14-18 April. <htpp://www7.scu.edu.au/programme/fullpapers/1875/com1875.htm>

Nissenbaum, H. 1998. Values in the Design of Computer Systems" *Computers in Society* March: 38-39.

Office of the Vice President, *Remarks as Delivered by Vice President Gore to the Networked Economy Conference*, 12 September 1995.

Phua, V. 1998. *Towards a set of ethical rules for search engines*. MSc dissertation, London Sc School of Economics.

Pollack, Andrew. 1995. A Cyberspace Front in a Multicultural War. *New York Times*. 7 August C1, C6.

Pollack, Andrew, and Hockley, A. 1997. What's wrong with internet searching. *D-Lib Magazine* March.

Poster, Mark. 1995. CyberDemocracy: Internet and the Public Sphere. In *Internet Culture*, ed. David Porter, pp. 201-217. New York: Routledge. <http://www.hnet.uci.edu/mposter/writings/democ.html>

Pringle, G., Allison, L., and Dowe, D.L. 1998. What is tall poppy among webpages. Paper presented at the *Seventh International World Wide Web Conference*, Brisbane, Australia, 14-18 April.<htpp://www7.scu.edu.au/programme/fullpapers/1872/com1872.htm>

Raboy, Marc. 1998. Global Communication Policy and Human Rights. *A Communications Cornucopia: Markle Foundation Essay's on Information Policy*, Washington DC: Brookings Institution Press.

Sen, Amartya. 1985. The Moral Standing of the Market. *Social Philosophy & Policy* 2:2 Spring.

Sen, Amartya. 1987. *On Ethics and Economics*. Oxford: Blackwell.

Shapiro, Andrew. 1995. Street Corners in Cyberspace. *The Nation* 3 July.

Schiller, Dan. 1995. Ambush on the I-Way: Information commoditization on the electronic frontier. *BCLA Information Policy Conference*, Vancouver, 27-28 October.

Sirbu, Marvin. 1992. Telecommunications Technology and Infrastructure. Institute for Information Studies. In *A national information network: Changing our lives in the 21st century,* pp. 174-175 Nashville, TN and Queenstown, MD: Institute for Information Studies.

Van den Hoven, Jeroen. 1994. Towards ethical principles for designing politico-administrative information systems. *Informatization in the Public Sector* 3: 353-373.

Wurman, R. S. 1989 *Information Anxiety: what to do when information doesn't tell you what you want to know,* New York: Bantam Books.