1-1-2006

SEARCH ENGINE BIAS AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

ERIC GOLDMAN

Follow this and additional works at: http://digitalcommons.law.yale.edu/yjolt

Part of the Computer Law Commons, Intellectual Property Commons, and the Science and Technology Commons

Recommended Citation

Available at: http://digitalcommons.law.yale.edu/yjolt/vol8/iss1/6

This Article is brought to you for free and open access by Yale Law School Legal Scholarship Repository. It has been accepted for inclusion in Yale Journal of Law and Technology by an authorized administrator of Yale Law School Legal Scholarship Repository. For more information, please contact julian.aiken@yale.edu.
SEARCH ENGINE BIAS
AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

ERIC GOLDMAN

ABSTRACT

Due to search engines’ automated operations, people often assume that search engines display search results neutrally and without bias. However, this perception is mistaken. Like any other media company, search engines affirmatively control their users’ experiences, which has the consequence of skewing search results (a phenomenon called “search engine bias”). Some commentators believe that search engine bias is a defect requiring legislative correction. Instead, this Essay argues that search engine bias is the beneficial consequence of search engines optimizing content for their users. The Essay further argues that the most problematic aspect of search engine bias, the “winner-take-all” effect caused by top placement in search results, will be mooted by emerging personalized search technology.

TABLE OF CONTENTS

I. SEARCH ENGINES MAKE EDITORIAL CHOICES .......................189
   A. INDEXING .........................................................................................190
   B. RANKING .......................................................................................191
II. SEARCH ENGINE EDITORIAL CHOICES CREATE BIASES .........192
III. SEARCH ENGINE BIAS IS NECESSARY AND DESIRABLE .........195
IV. TECHNOLOGICAL EVOLUTION WILL MOOT SEARCH ENGINE BIAS 198
V. CONCLUSION ..................................................................................199

* Assistant Professor, Marquette University Law School, Milwaukee, Wisconsin. Home page: http://www.ericgoldman.org. Email: egoldman@gmail.com. I appreciate the comments of participants at the Yale Law School Regulating Search Symposium and the 2005 Association of Internet Researchers (AoIR) Annual Meeting in Chicago.
SEARCH ENGINE BIAS AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

In the past few years, search engines have emerged as a major force in our information economy, helping searchers perform hundreds of millions (or even billions) of searches per day.¹ With this broad reach, search engines have significant power to shape searcher behavior and perceptions. In turn, the choices that search engines make about how to collect and present data can have significant social implications.

Typically, search engines automate their core operations, including the processes that search engines use to aggregate their databases and then sort/rank the data for presentation to searchers. This automation gives search engines a veneer of objectivity and credibility.² Machines, not humans, appear to make the crucial judgments, creating the impression that search engines bypass the structural biases and skewed data presentations inherent in any human-edited media.³

Unfortunately, this romanticized view of search engines does not match reality. Search engines are media companies. Like other media companies, search engines make editorial choices designed to satisfy their audience.⁴ These choices systematically favor certain types of content over others, producing a phenomenon called “search engine bias.”

Search engine bias sounds scary, but this Essay explains why such bias is both necessary and desirable. The Essay also explains how emerging personalization technology will soon ameliorate many concerns about search engine bias.

I. SEARCH ENGINES MAKE EDITORIAL CHOICES

Search engines frequently claim that their core operations are completely automated and free from human intervention,⁵ but this

⁴ See, e.g., C. EDWIN BAKER, ADVERTISING AND A DEMOCRATIC PRESS (1994).
characterization is false. Instead, humans make numerous editorial judgments about what data to collect and how to present that data.

A. INDEXING

Search engines do not index every scrap of data available on the Internet. Search engines omit (deliberately or accidentally) some web pages entirely or may incorporate only part of a web page. During indexing, search engines are designed to associate third party “metadata” (data about data) with the indexed web page. For example, search engines may use and display third party descriptions of the website in the search results. Search engines may also index “anchor text” (the text that third parties use in hyperlinking to a website), which can cause a website to appear in search results for a term the website never used (and may object to).

not censor results for any search terms. The order and content of our results are completely automated; we do not manipulate our search results by hand.”; Technology Overview, Google.com, http://www.google.com/corporate/tech.html (“There is no human involvement or manipulation of results....”); see also Complaint at ¶¶ 37-38, 52-56, KinderStart.com LLC v. Google, Inc., Case No. C 06-2057 RS (N.D. Cal. Mar. 17, 2006) (giving other examples of Google’s claims to be passive). Note that Google has subsequently revised some of these cited pages after its censorship controversy in China.


10 For example, the first search result in Google and Yahoo! for the keyword “miserable failure” is President George W. Bush’s home page because so many websites have linked to the biography using the term “miserable failure.” See Tom McNichol, Your Message Here, N.Y. TIMES, Jan. 22, 2004, at G1. This algorithmic vulnerability has spawned a phenomenon called “Google bombing,” where websites coordinate an anchor text attack to
Finally, once indexed, search engines may choose to exclude web pages from their indexes for a variety of reasons, ranging from violations of quasi-objective search engine technical requirements to simple capriciousness.12

B. RANKING

To determine the order of search results, search engines use complex proprietary “ranking algorithms.” Ranking algorithms obviate the need for humans to make individualized ranking decisions for the millions of search terms used by searchers, but they do not lessen the role of human editorial judgment in the process. Instead, the choice of which factors to include in the ranking algorithm, and how to weight them, reflects the search engine operator’s editorial judgments about what makes content valuable. Indeed, to ensure that these judgments are produce desired results, search engines manually inspect search results13 and make adjustments accordingly.

Additionally, search engines claim they do not modify algorithmically-generated search results, but there is some evidence to the contrary. Search engines allegedly make manual adjustments of a web publisher’s overall ranking,14 and search engines occasionally modify search results presented in response to particular keyword searches.’


12 This is the heart of KinderStart’s allegations against Google. See Complaint, KinderStart.com LLC v. Google, Inc., Case No. C 06-2057 (N.D. Cal. Mar. 17, 2006). Although the complaint’s allegations about Google’s core algorithmic search may not be proven, Google does liberally excise sources from Google News. For example, Google claims that “news sources are selected without regard to political viewpoint or ideology,” see Google News (Beta), Google.com, http://news.google.com/intl/en_us/about_google_news.html#25, but Google dropped a white supremacist news source from Google News because it allegedly promulgated “hate content.” See Susan Kuchinskas, Google Axes Hate News, INTERNETNEWS.COM, Mar. 23, 2005, http://www.internetnews.com/xSP/article.php/3492361.


http://digitalcommons.law.yale.edu/yjolt/vol8/iss1/6
Conclusion

Search engines have some duality in their self-perceptions, and this duality creates a lot of confusion. Search engines perceive themselves as objective and neutral because they let automated technology do most of the hard work. However, in practice, search engines make editorial judgments just like any other media company. Principally, these editorial judgments are instantiated in the parameters set for the automated operations, but search engines also make individualized judgments about what data to collect and how to present it. These manual interventions may be the exception and not the rule, but these exceptions only reinforce that search engines play an active role in shaping their users’ experiences when necessary to accomplish their editorial goals.

II. SEARCH ENGINE EDITORIAL CHOICES CREATE BIASES

Search results ordering has a significant effect on searchers and web publishers. Searchers usually consider only the top few search results; the top-ranked search result gets a high percentage of searcher clicks, and clickthrough rates quickly decline from there. Therefore, even if a search

blocked certain search terms containing the keyword “phpBB”). Other examples of search engines manually adjusting algorithmically-generated results:

- In response to the search term “Jew,” for a period of time (including, at minimum November 2005 when the author observed the phenomenon), Google displayed a special result in the sponsored link, saying “Offensive Search Results: We’re disturbed about these results as well. Please read our note here.” The link led to a page explaining the results (see http://www.google.com/explanation.html).
- Amazon’s book search functionality offered the prompt “did you mean adoption?” in response to searches for the keyword “abortion.” When pointed out to Amazon, it manually changed the database to disable that prompt. See Laurie J. Flynn, Amazon Says Technology, Not Ideology, Skewed Results, N.Y. TIMES, Mar. 20, 2006, at B8.
- Google removed some websites from its index in response to a 512(c)(3) take-down demand from the Church of Scientology. However, Google displayed the following legend at the bottom of affected search results pages (such as search results for “scientology site:xenu.net”): “In response to a complaint we received under the US Digital Millennium Copyright Act, we have removed 2 result(s) from this page. If you wish, you may read the DMCA complaint that caused the removal(s) at ChillingEffects.org.” See http://www.google.com/search?sourceid=navclient&ie=UTF-8&rls=GGLD,GGLD:2005-09,GGLD:cn&q=scientology+site%3Axenu%2Enet (go to google.com, enter “scientology.site.xenu.net”, then click search and scroll to the bottom of the page); see also Chris Sherman, Google Makes Scientology Infringement Demand Public, SEARCH ENGINE WATCH, Apr. 15, 2002, http://searchenginewatch.com/searchday/article.php/2159691.

---


Yale Journal of Law and Technology Spring 2006

Published by Yale Law School Legal Scholarship Repository, 2006
SEARCH ENGINE BIAS AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

engine delivers hundreds or even thousands of search results in response to a searcher’s query, searchers effectively ignore the vast majority of those search results. Accordingly, web publishers desperately want to be listed among the top few search results.17

For search engines, results placement determines how the searcher perceives the search experience. If the top few search results do not satisfy the searcher’s objectives, the searcher may deem the search a failure. Therefore, to maximize searcher perceptions of search success, search engines generally tune their ranking algorithms to support majority interests.18 In turn, minority interests (and the websites catering to them) often receive marginal exposure in search results.

To gauge majority interests, search engines frequently include a popularity metric in their ranking algorithm. Google’s popularity metric, PageRank, treats inbound links to a website as popularity votes, but votes are not counted equally; links from more popular websites count more than links from lesser-known websites.19

Beyond promoting search results designed to satisfy majority interests, PageRank’s non-egalitarian voting structure causes search results to be biased towards websites with economic power20 because these websites get lots of links due to their marketing expenditures and general prominence.

Indeed, popularity-based ranking algorithms may reinforce and perpetuate existing power structures.21 Websites that are part of the current power elite get better search result placement, which leads to greater

Joachims that the first search result gets 42% of clicks and the second search result gets 8%; further, when the first two search results are switched, the first search result gets 34%—meaning that positioning dictated searcher behavior); Nico Brooks, The Atlas Rank Report: How Search Engine Rank Impacts Traffic, ATLAS INSTITUTE DIGITAL MARKETING INSIGHTS, June 2004, http://app.atlasonepoint.com/pdf/AtlasRankReport.pdf (the first ranked search result may get ten times the quantity of clicks as the tenth ranked search result).


21 See Introna & Nissenbaum, supra note 18.
consideration of their messages and views. Furthermore, the increased exposure attributable to better placement means that these websites are likely to get more votes in the future, leading to a self-reinforcing process. In contrast, minority-interest and disenfranchised websites may have a difficult time cracking through the popularity contest, potentially leaving them perpetually relegated to the search results hinterlands.

A number of commentators have lamented these effects and offered some proposals in response:

1. Improve Search Engine Transparency

Search engines keep their ranking algorithms secret. This secrecy hinders search engine spammers from gaining more prominence than search engines want them to have, but the secrecy also prevents searchers and commentators from accurately assessing any bias. To enlighten searchers, search engines could be required to disclose more about their practices and their algorithms. This additional information has two putative benefits. First, it may improve market mechanisms by helping searchers choose among search engine competitors. Second, it may help searchers determine the appropriate level of cognitive authority to assign to their search results.

2. Publicly Fund Search Engines

Arguably, search engines have “public good”-like attributes, such as reducing the social costs of search behavior. If so, private actors will not incorporate these social benefits into their decision-making. In that case, public funding of search engines may be required to produce socially

---


23 See Cho & Roy, supra note 22.


25 See Introna & Nissenbaum, supra note 18.
SEARCH ENGINE BIAS AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

optimal search results. Indeed, there have been several proposals to create government-funded search engines.

3. Mandate Changes to Ranking/Sorting Practices

Search engines could be forced to increase the exposure of otherwise-marginalized websites. At least two lawsuits (the Search King and KinderStart lawsuits against Google) have sought this result. In addition, some academics support mandatory reordering of search results. For example, Pandey et al. advocate a “randomized rank promotion” scheme where obscure websites randomly should get extra credit in ranking algorithms, appearing higher in the search results on occasion and getting additional exposure to searchers accordingly. As another example, Pasquale proposes that, when people think the search engines are providing false or misleading information, search engines should be forced to include a link to corrective information.

III. SEARCH ENGINE BIAS IS NECESSARY AND DESIRABLE

Before trying to solve the problem of search engine bias, we should be clear how search engine bias creates a problem that requires correction. From my perspective, search engine bias is the unavoidable consequence of search engines exercising editorial control over their databases. Like any other media company, search engines simply cannot passively and neutrally redistribute third party content (in this case, web publisher content). If a search engine does not attempt to organize web content, its system quickly

28 As Google said in its response to the KinderStart lawsuit, “Plaintiff KinderStart contends that the judiciary should have the final say over [search engines’] editorial process. It has brought this litigation in the hopes that the Court will second-guess Google’s search rankings and order Google to view KinderStart’s site more favorably.” Motion to Dismiss at 1, KinderStart.com LLC v. Google, Inc., No. C 06-2057 RS (N.D. Cal. May 2, 2006).
29 See Sandeep Pandey et al., Shuffling a Stacked Deck: the Case for Partially Randomized Ranking of Search Engine Results, http://www.cs.cmu.edu/~olston/publications/randomRanking.pdf; cf. SUNSTEIN, supra note 26 (explaining that websites should be forced to link to contrary views as a way of increasing exposure to alternative viewpoints).
30 See Pasquale, supra note 20, at 28-30 (proposing that the link be displayed as an asterisk to the search results).
and inevitably will be overtaken by spammers, fraudsters and malcontents.\textsuperscript{31}
At that point, the search engine become worthless to searchers.

To prevent anarchy and preserve credibility, search engines unavoidably must exercise some editorial control over their systems. In turn, this editorial control will create some bias.

Fortunately, market forces limit the scope of search engine bias. Searchers have high expectations for search engines: they expect search engines to read their minds\textsuperscript{32} and infer their intent based solely on a small number of search keywords.\textsuperscript{33} Search engines that disappoint (either by failing to deliver relevant results, or by burying relevant results under too many unhelpful results) are accountable to fickle searchers.\textsuperscript{34} There are

\textsuperscript{31} Every Internet venue accepting user-submitted content inevitably gets attacked by unwanted content. If left untended, the venue inexorably degrades into anarchy. See, e.g., Alorie Gilbert, Google Fixes Glitch That Unleashed Flood of Porn, CNET News.com, Nov. 28, 2005, http://news.com.com/2102-1025_3-5969799.html?tag=st.util.print (describing how Google Base, a venue for user-submitted content, was overtaken by pornographers: "the amount of adult content on Google Base was staggering considering Google only launched the tool a week ago."); Josh Quittner, The War Between alt.tasteless and rec.pets.cats, WIRED, May 1994, at 46 (describing how a group of anarchists, for fun, took over a USENET newsgroup about pets).


\textsuperscript{33} Searchers routinely use a very small number of keywords to express their search interests. See iProspect.com, Inc., iProspect Natural SEO Keyword Length Study, Nov. 2004, http://www.iprospect.com/premiumPDFs/keyword_length_study.pdf (eighty-eight percent of search engine referrals are based on only one or two keywords); see also Declan Butler, Souped-Up Search Engines, NATURE, May 11, 2000, at 112, 115 (citing an NEC Research Institute study showing that up to 70% of searchers use only a single keyword as a search term); Bernard J. Jansen et al., Real Life Information Retrieval: A Study of User Queries on the Web, 32 SIGIR FORUM 5, 15 (1998) (stating that the average keyword length was 2.35 words; one-third of searches used one keyword and 80% used three keywords or fewer); Jakob Nielsen, JAKOB NIELSEN'S ALERTBOX, Search: Visible and Simple, May 13, 2001, http://www.useit.com/alertbox/20010513.html (stating that the average keyword length was 2.0 words).

\textsuperscript{34} See Kim Peterson, Microsoft Learns to Crawl, SEATTLE TIMES, May 2, 2005 (MSN Search "learned that the arcane searches were the make-or-break moments for Web searchers. People weren't just happy when a search engine could find answers to their most bizarre, obscure and difficult queries. They would switch loyalties."); Bob Tedeschi, Every Click You Make, They'll Be Watching You, N.Y. TIMES, Apr. 3, 2006, http://www.nytimes.com/2006/04/03/business/03ecom.html?ei=5090&en=9e55ac64f692433a&ex=1301716800&partner=rssuserland&emc=rss&pagewanted=print.
SEARCH ENGINE BIAS AND THE DEMISE OF SEARCH ENGINE UTOPIANISM

multiple search engines available to searchers, and few barriers to switching between them.

As a result, searchers will shop around if they do not get the results they want, and this competitive pressure constrains search engine bias. If a search engine’s bias degrades the relevancy of search results, searchers will explore alternatives even if searchers do not realize that the results are biased. Meanwhile, search engine proliferation means that niche search engines can segment the market and cater to underserved minority interests. Admittedly, these market forces are incomplete—searchers may never consider what results they are not seeing—but they are powerful nonetheless.

In contrast, it is hard to imagine how regulatory intervention will improve the situation. First, regulatory solutions become a vehicle for normative views about what searchers should see—or should want to see. How should we select among these normative views? What makes one bias better than the other?

Second, regulatory intervention that promotes some search results over others does not ensure that searchers will find the promoted search results useful. Instead, government regulation rarely can do better than

35 In addition to the recent launch of major new search engines by providers like MSN, the open-source software community is developing Nutch to allow anyone to build and customize his or her own web search engine. http://lucene.apache.org/nutch/; see also Olsen, Open-Source Niche, supra note 24.
36 See Rahul Telang et al., An Empirical Analysis of Internet Search Engine Choice, Aug. 2002. On file with author. For example, search engines use the same basic interface (a white search box), and searchers rarely use advanced search features that might require additional learning time at other search engines.
39 See, e.g., Susan L. Gerhart, Do Web Search Engines Suppress Controversy?, FIRST MONDAY, Jan. 2004, http://www.firstmonday.org/issues/issue9_1/gerhart/. Gerhart argues that search engines do not adequately prioritize search results that expose controversies about the search topic. However, her argument assumes that controversy-related information has value to consumers, an assumption that deserves careful evaluation.
market forces at delivering results that searchers find relevant, so searchers likely will find some of the promoted results irrelevant.

The clutter of unhelpful result may hinder searchers’ ability to satisfy their search objectives, undermining searchers’ confidence in search engines’ mind-reading abilities. In this case, regulatory intervention could counterproductively degrade search engines’ value to searchers. Whatever the adverse consequences of search engine bias, the consequences of regulatory correction are probably worse.

IV. TECHNOLOGICAL EVOLUTION WILL MOOT SEARCH ENGINE BIAS

Currently, search engines use “one-size-fits-all” ranking algorithms to deliver homogeneous search results to searchers with heterogeneous search objectives. One-size-fits-all algorithms exacerbate the consequences of search engine bias in two ways: (1) it creates winners (websites listed high in the search results) and losers (those with marginal placement), and (2) it delivers suboptimal results for searchers with minority interests.

These consequences will abate when search engines migrate away from one-size-fits-all algorithms towards “personalized” ranking algorithms. Personalized algorithms produce search results that are custom-tailored to each searcher’s interests, so searchers will see different results in response to the same search query. For example, Google offers searchers an option that “orders your search results based on your past searches, as well as the search results and news headlines you’ve clicked on.”

Personalized ranking algorithms represent the next major advance in search relevancy. One-size-fits-all ranking algorithms have inherent limits on their maximum relevancy potential, and further improvements in one-

---

44 See Pitkow, supra note 42, at 50.
size-fits algorithms will yield progressively smaller relevancy benefits. Personalized algorithms transcend those limits, optimizing relevancy for each searcher and thus implicitly doing a better job of searcher mind-reading.  

Personalized ranking algorithms also reduce the effects of search engine bias. Personalized algorithms mean that there are multiple “top” search results for a particular search term instead of a single “winner,” so web publishers will not compete against each other in a zero-sum game. Also, personalized algorithms necessarily will diminish the weight given to popularity-based metrics (to give more weight for searcher-specific factors), reducing the structural biases due to popularity. Personalized ranking algorithms are not a panacea—any process where humans select and weight algorithmic factors will produce some bias—but personalized algorithms will eliminate many of the current concerns about search engine bias.

V. CONCLUSION

Complaints about search engine bias implicitly reflect some disappointed expectations. In theory, search engines can transcend the deficiencies of predecessor media to produce a type of media utopia. In practice, search engines are just like every other medium—heavily reliant on editorial control and susceptible to human biases. This fact shatters any illusions of search engine utopianism.

Fortunately, search engine bias may be largely temporal. In this respect, I see strong parallels between search engine bias and the late 1990s keyword metatag “problem.” Web publishers used keyword metatags to distort search results, but these techniques worked only so long as search engines considered keyword metatags in their ranking algorithms. When search engines recognized the distortive effects of keyword metatags, they changed their algorithms to ignore keyword metatags. Search result relevancy improved, and the problem was solved without regulatory intervention.

46 See Jaime Teevan et al., Personalizing Search via Automated Analysis of Interests and Activities, SIGIR ’05, http://haystack.lcs.mit.edu/papers/teevan.sigir05.pdf; Terry McCarthy, On the Frontier of Search, TIME, Aug. 28, 2005 (“Search will ultimately be as good as having 1,000 human experts who know your tastes scanning billions of documents within a split second.”) (quoting Gary Flake, Microsoft Distinguished Engineer).


48 Personalized algorithms have other potentially adverse consequences, such as creating self-reinforcing information flows. See SUNSTEIN, supra note 26. For a critique of these consequences, see Goldman, Coasean Analysis, supra note 40.

49 See generally Goldman, Deregulating Relevancy, supra note 7.

Similarly, search engines naturally will continue to evolve their ranking algorithms and improve search result relevancy—a process that, organically, will cause the most problematic aspects of search engine bias to largely disappear. To avoid undercutting search engines’ quest for relevance, this effort should proceed without regulatory distortion.