DESIGNING AN ETHICAL SEARCH ENGINE: THE INTERPLAY BETWEEN INDEX AND RANKING

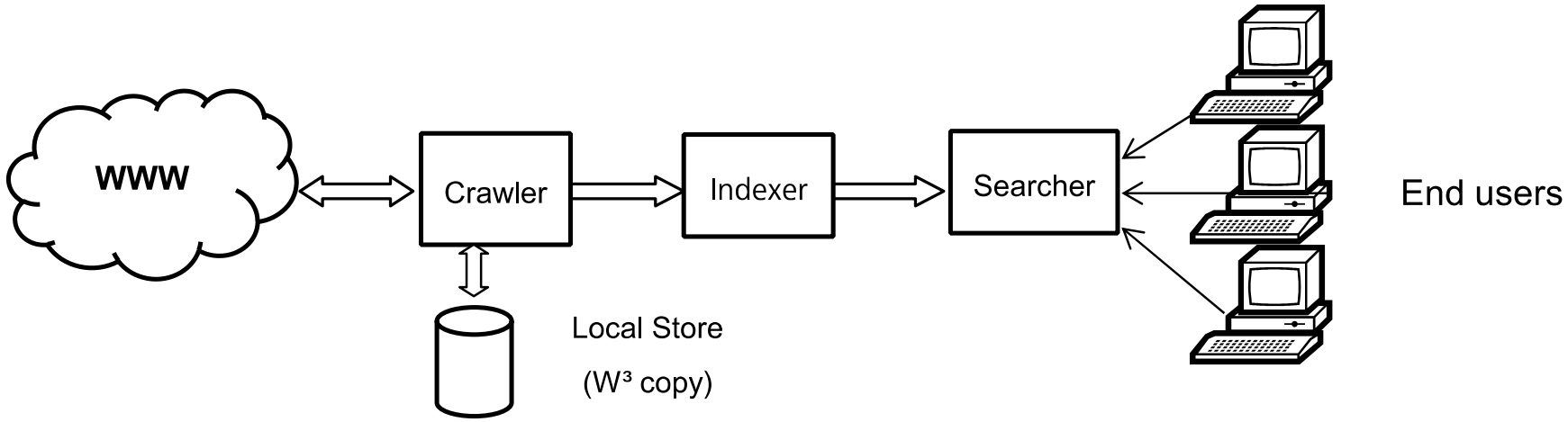
Workshop „Search Engine Ethics”

University of Birmingham, 13 September 2024

Dirk Lewandowski
Hamburg University of Applied Sciences, Hamburg, Germany
University of Duisburg-Essen, Duisburg, Germany
dirk.lewandowski@haw-hamburg.de

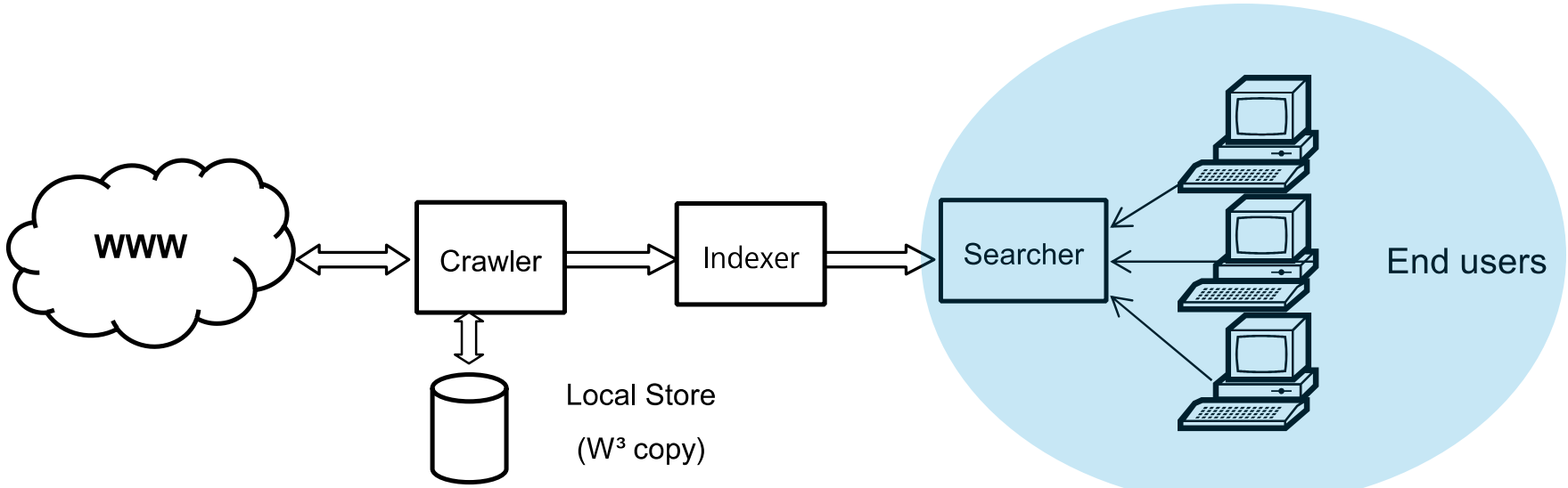
HAW HAMBURG

SEARCH ENGINE COMPONENTS



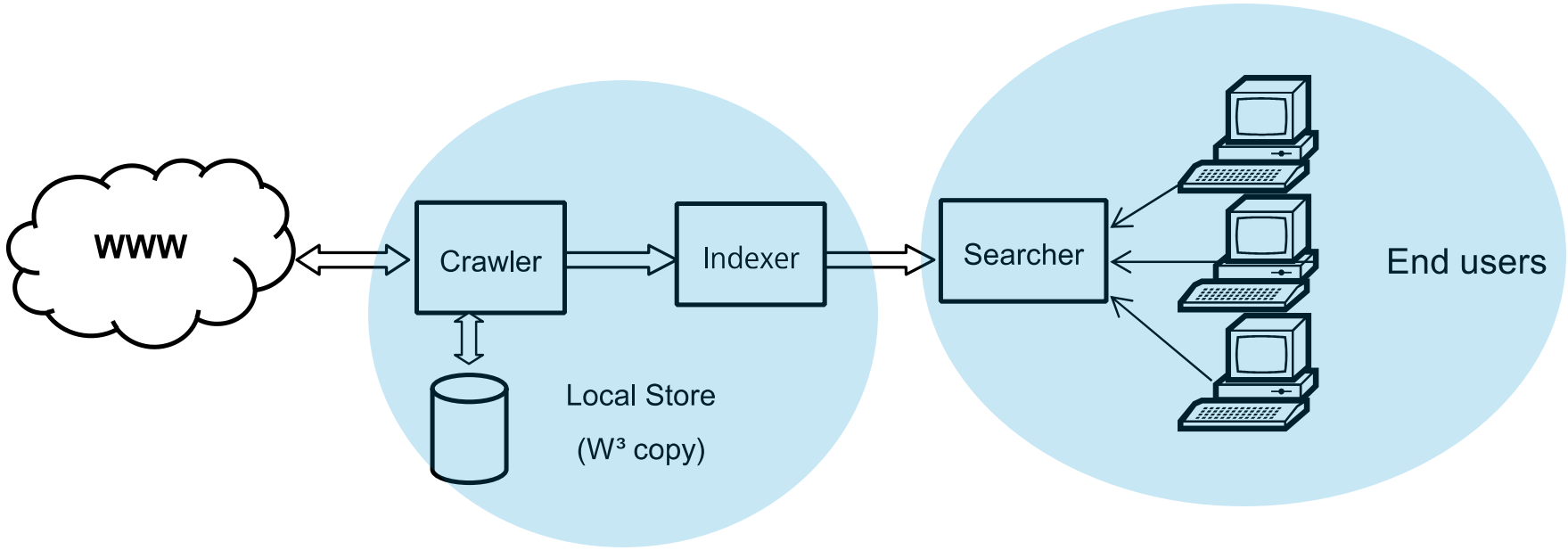
(from Risvik & Michelsen, 2002)

SEARCH ENGINE COMPONENTS



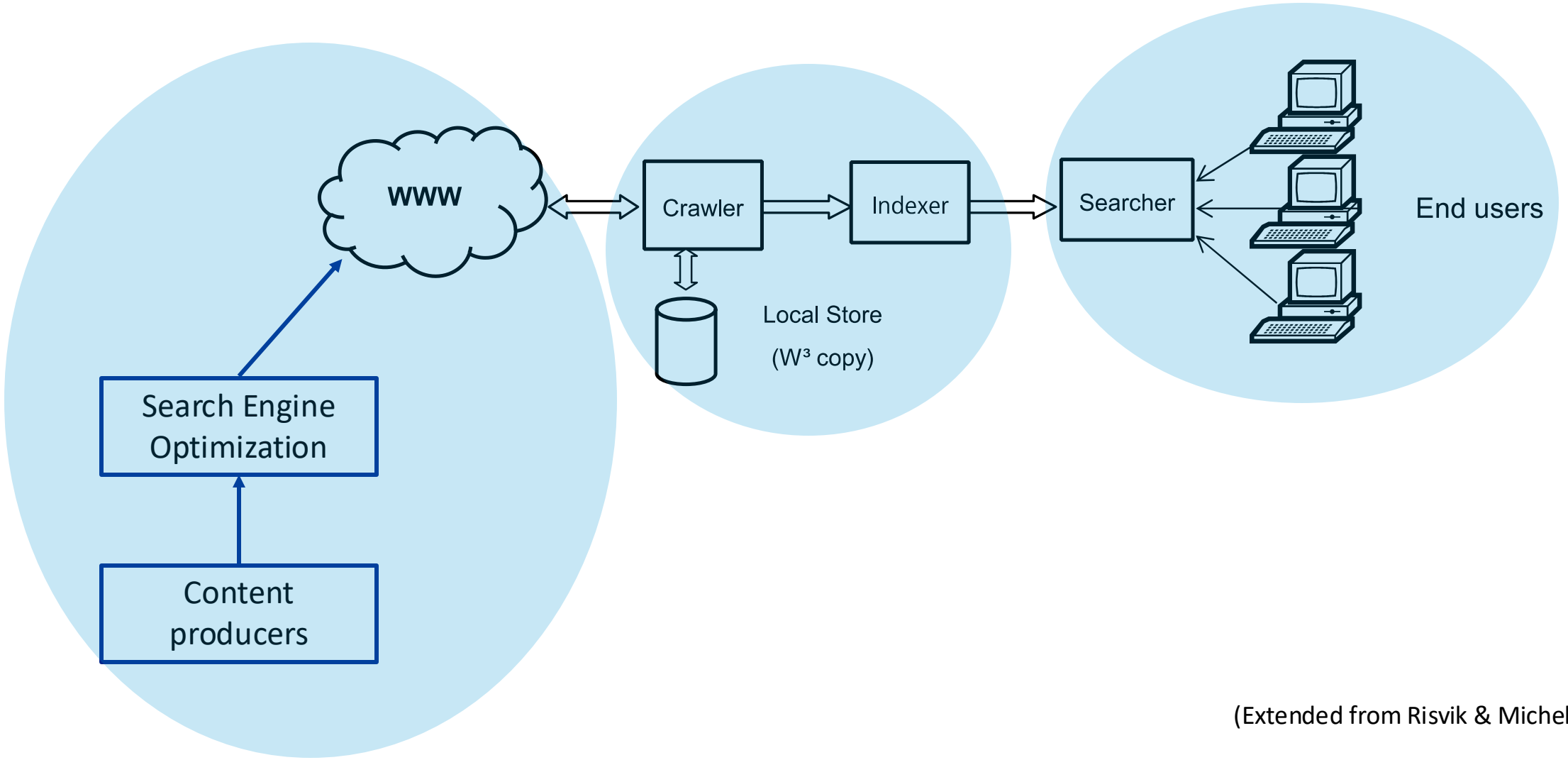
(Extended from Risvik & Michelsen, 2002)

SEARCH ENGINE COMPONENTS



(Extended from Risvik & Michelsen, 2002)

SEARCH ENGINE COMPONENTS



(Extended from Risvik & Michelsen, 2002)

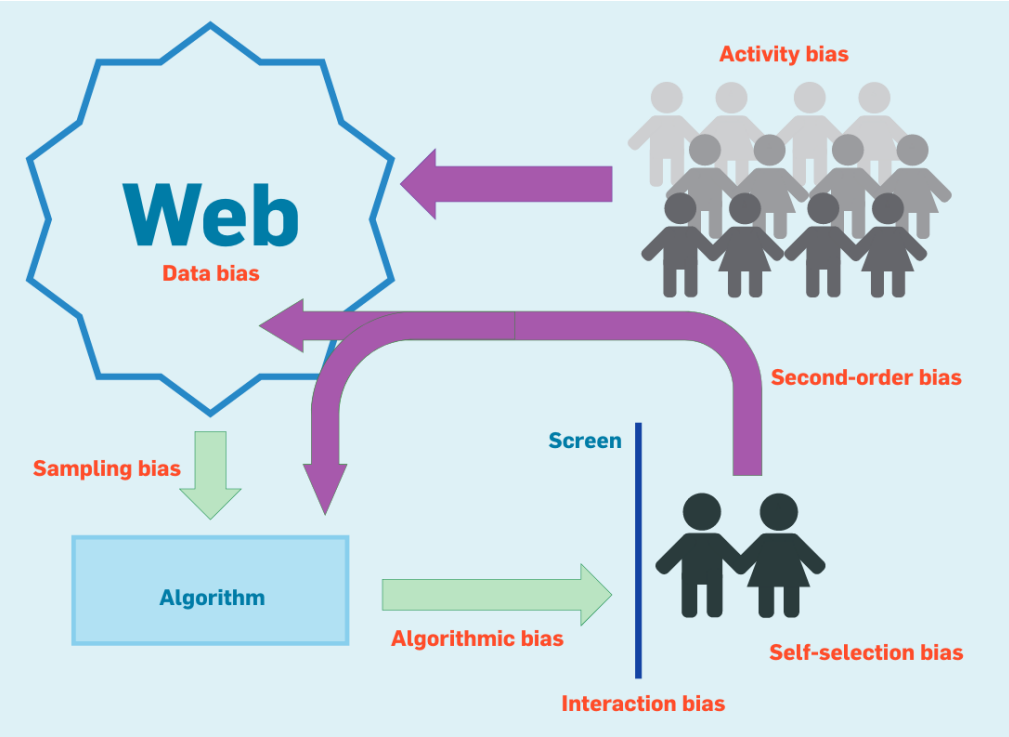
INCLUSION, EXCLUSION AND CURATION

General approach to content acquisition

- Search engines try to index as much of the web as possible (although financial and technical resources limit them).
- Controlling the index focuses on excluding spam (or very low-quality content rather than making active choices about which sources to include.
- A problem resulting from this approach is that everything in the index can show up, at least for some obscure query.
- If the index is used to train AI models, that content will also influence the AI-generated answers.

What is produced?

- Some (types of) content is produced in vast amounts, mainly for commercial reasons (“data bias”, see Baeza-Yates, 2018).
- No current search engine limits itself to index only selected sources (“curation”).



(Baeza-Yates, 2018, p. 56)

PROBLEM/SITUATION

Problems resulting from “indexing everything”

- Search engines represent the flood of misinformation and disinformation.
- Search engines’ answer to misinformation and low-quality content is to downrank it in the hopes that users will not see it, or at least only when entering very specific queries.
- Still based on the assumption that an ideal search engine index should be a complete and current “copy of the web”.

Building „ethical search engines“ on content curation

- Search is only in part about ranking. It is also about collecting and representing content.
- Focusing on the interplay between the index and ranking function goes beyond just mitigating bias through ranking.
- “Ethical search engines” need to curate their indexes.

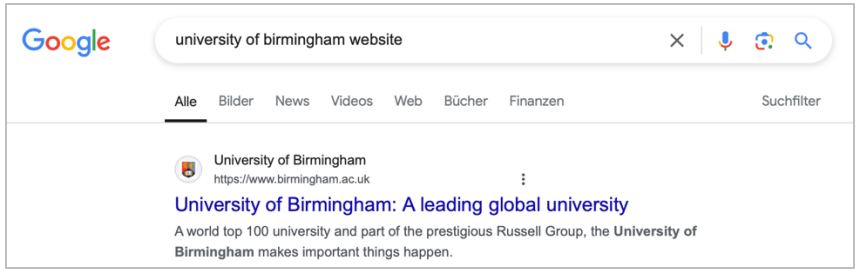
Document lists vs (LLM-generated) answers

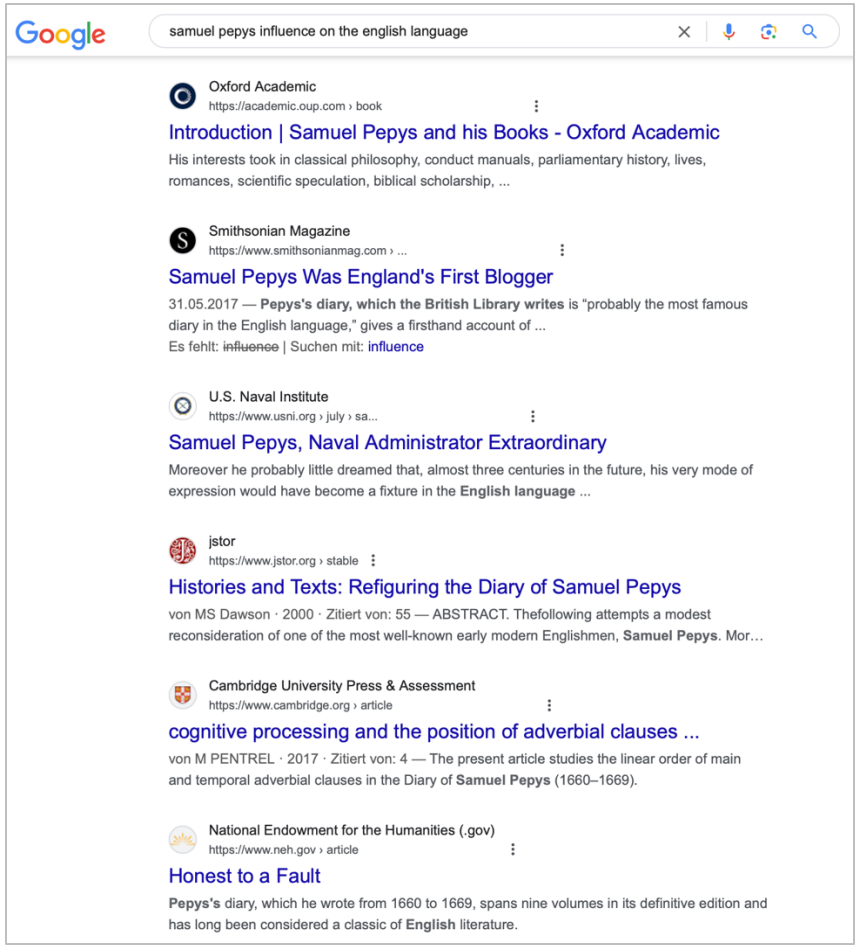
- The following discussion applies to both generating lists of documents in response to a query (the “traditional” search engine approach) and generating complex responses using generative artificial intelligence.
- While the technical basis for generating search responses changes with AI, the more fundamental problem of supporting users with information to fulfil their information needs or solve their problems remains the same.

RANKING AND ALGORITHMIC INTERPRETATION

Every ranking is an interpretation of the content in the index

- An unbiased ranking is possible only for some queries/query types: navigational queries, factual information.
- As every search engine is biased as it interprets the web's content, an unbiased search engine is simply impossible (Lewandowski, 2017).
- "Search engines always provide us with a version of reality. Thus, what is interesting and relevant is to map more precisely which or whose values are represented and how." (Haider & Sundin, 2019, p. 31)
- This leads to how to represent the content collected in the index in the ranking function.
- A ranking function can be seen as a function building a result set from a document set, meaning selecting some documents over others to generate answers.
- Applying formal criteria under the assumption they do not inherit preferences/judgments is misleading.





THE NEED FOR COMPLEX INTERPRETATIONS

Three ranking approaches based on external formal criteria (*simplified*), leading to a "neutral" representation of the results in the index.

Representing the content from the Web proportionally

- Would merely reflect the biases of the web index ("data bias").
- The proportion of web content is not a good indicator of a claim's actual importance or truthfulness (e.g., in health symptom explanations, see White & Horvitz, 2009).

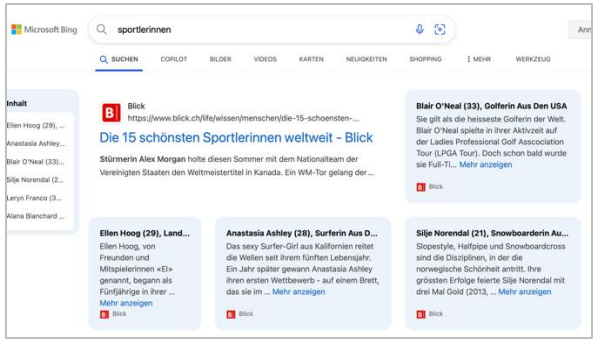
Table I. Probability of Mention of Cause Given Symptom

Symptom	Cause	Web Crawl	Web Search	Domain Search
headache	caffeine withdrawal	.29	.26	.25
	tension	.68	.48	.75
	brain tumor	.03	.26	.00
muscle twitches	benign fasciculation	.53	.12	.34
	muscle strain	.40	.38	.66
	ALS	.07	.50	.00
chest pain	indigestion	.28	.35	.38
	heartburn	.57	.28	.52
	heart attack	.15	.37	.10

(White & Horvitz, 2009)

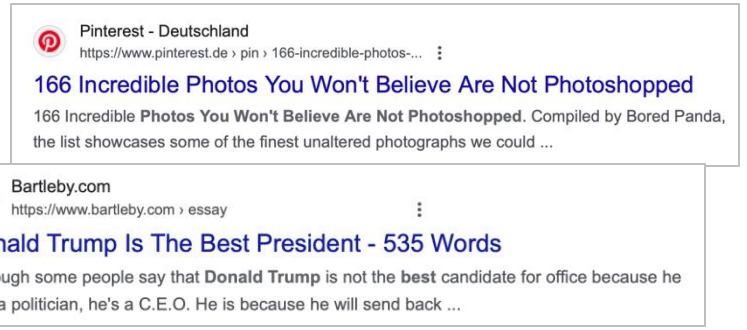
Representation of user interest

- Focusing on meaningful user click-through rates (CTR) would reproduce the preferences and biases of the user population.
- Any minority position would be ranked low, leading to few users even considering them.



Personalisation of results

- Would lead to individual users getting the results they would be most happy with without considering opposing views, different perspectives, etc.
- Would lead to negative consequences like the "filter bubble" (Pariser, 2011)



SOCIETAL RELEVANCE

How to model relevance in an ethical fashion

- Haider and Sundin (2019) proposed the concept of societal relevance as an addition to the classic distinction between system and user relevance in information science (Saracevic, 2016).
- The idea has not yet been fully developed, and it is unclear when some content should be downranked and when it should be excluded from the index.
- What part of it actually is" relevance" (a relation to sth.), and which is information quality (a property of the information object)?
- If information quality, a matter of curating the index.

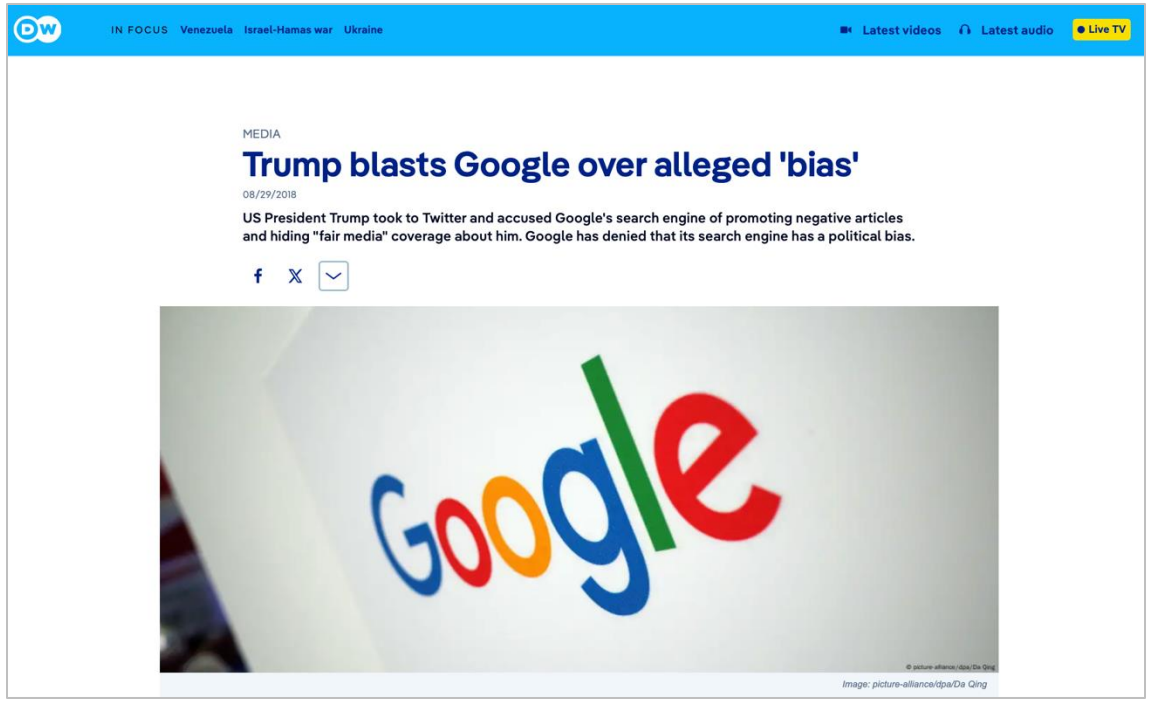
Does "societal relevance" depend on the particular query?

- For what kind of queries should particular results/answers be shown?
- "What did Donald Trump say about how you can treat women when you're rich and famous?" vs "How should you treat women?"

ETHICAL AND UNETHICAL SEARCH ENGINES

“Unethical” search engines

- Based on the “algorithmic ideology” underlying current search engines (e.g., Mager, 2012)?
- Based on showing discriminating results in top positions (most prominently, Noble, 2018)?
- It remains unclear on which grounds results should *not* be displayed or downranked.
- Argument: Bias + people feeling discriminated by search results.
- The problem is reliably measuring bias or discrimination in search results.



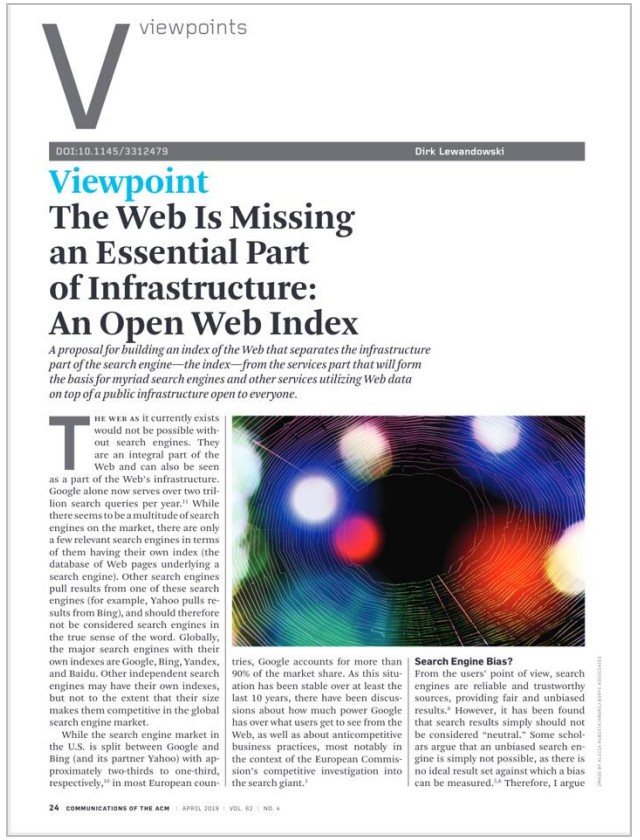
DESIGNING AN ”ETHICAL SEARCH ENGINE”

Properties of an ”ethical search engine”

- Based on values, deriving *index curation and ranking* from these values.
- Only one of many possible interpretations of the web’s content.
- No complete representation of web content but will deliberately exclude some sources deemed unethical.
- Calls for transparency about curation (i.e., what this search engine includes in its index).

More than one ”ethical search engine”

- Different values and/or a different weighting of these values will lead to different algorithmic interpretations (search results).
- This calls for not one more search engine but many new search engines.
- Given the barriers to entering the search engine market (Lewandowski, 2023, chapter 8), the solution can only lie in building an Open Web Index (OWI) as a public infrastructure that allows any search engine to develop its services on top (Lewandowski, 2019).



viewpoints

DOI:10.1145/3312479 Dirk Lewandowski

Viewpoint

The Web Is Missing an Essential Part of Infrastructure: An Open Web Index

A proposal for building an index of the Web that separates the infrastructure part of the search engine—the index—from the services part that will form the basis for myriad search engines and other services utilizing Web data on top of a public infrastructure open to everyone.

THE WEB AS it currently exists would not be possible without search engines. They are an integral part of the Web and can also be seen as a part of the Web's infrastructure. Google alone now serves over two trillion search queries per year.[1] While there seems to be a multitude of search engines on the market, there are only a few relevant search engines in terms of them having their own index (the database of Web pages underlying a search engine). Other search engines pull results from one of these search engines (for example, Yahoo pulls results from Bing), and should therefore not be considered search engines in the true sense of the word. Globally, the major search engines with their own indexes are Google, Bing, Yandex, and Baidu. Other independent search engines may have their own indexes, but not to the extent that their size makes them competitive in the global search engine market.

While the search engine market in the U.S. is split between Google and Bing (and its partner Yahoo) with approximately two-thirds to one-third, respectively,[2] in most European countries, Google accounts for more than 90% of the market share. As this situation has been stable over at least the last 10 years, there have been discussions about how much power Google has over what users get to see from the Web, as well as about anticompetitive business practices, most notably in the context of the European Commission's competitive investigation into the search giant.[3]

Search Engine Bias?

From the users' point of view, search engines are reliable and trustworthy sources, providing fair and unbiased results.[4] However, it has been found that search results simply should not be considered "neutral." Some scholars argue that an unbiased search engine is simply not possible, as there is no ideal result set against which a bias can be measured.[5,6] Therefore, I argue

24 COMMUNICATIONS OF THE ACM | APRIL 2019 | VOL. 62 | NO. 4

REFERENCES

HAW HAMBURG

REFERENCES

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Haider, J., & Sundin, O. (2019). *Invisible Search and Online Search Engines*. Routledge.

Lewandowski, D. (2017). Is Google Responsible for Providing Fair and Unbiased Results? In M. Taddeo & L. Floridi (Eds.), *The Responsibilities of Online Service Providers* (Vol. 31, pp. 61–77). Springer. https://doi.org/10.1007/978-3-319-47852-4_4

Lewandowski, D. (2019). The web is missing an essential part of infrastructure: An Open Web Index. *Communications of the ACM*, *62*(4), 24–27. https://doi.org/10.1145/3312479

Lewandowski, D. (2023). *Understanding Search Engines*. Springer International Publishing. https://doi.org/10.1007/978-3-031-22789-9

Mager, A. (2012). Algorithmic Ideology: How capitalist society shapes search engines. *Information, Communication & Society*, *15*(5), 1–19. https://doi.org/10.1080/1369118X.2012.676056

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Viking.

Risvik, K. M., & Michelsen, R. (2002). Search engines and web dynamics. *Computer Networks*, *39*(3), 289–302.

Saracevic, T. (2016). The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *8*(3), i–109. https://doi.org/10.2200/S00723ED1V01Y201607ICR050

White, R. W., & Horvitz, E. (2009). Cyberchondria. *ACM Transactions on Information Systems*, *27*(4), Article No. 23. https://doi.org/10.1145/1629096.1629101