Towards a Framework for Harm Prevention in Web Search

Steven Zimmerman[1], Stefan M. Herzog[2], David Elsweiler[3], Jon Chamberlain[1], and Udo Kruschwitz[3]

[1] University of Essex, Colchester, United Kingdom
[2] Center for Adaptive Rationality Max Planck Institute for Human Development, Berlin, Germany
[3] Universität Regensburg, Regensburg, Germany

Abstract. We introduce a framework aimed at the information science (IS), information retrieval (IR) and data science communities as well as behavioral and cognitive scientists and policy makers inside government and at corporations operating Web search platforms. The goal of this framework is to instigate collaborative discussion and research across these communities to address potential dangers searchers and society face in modern Web search. We provide an overview of the harms, such as poor health outcomes, and their possible causes, including searcher and system biases being tuned for maximum profit. Modifications to policy, additional evaluation metrics, a mixture of cognitive decision making tools and improvements to the IR system are the suggested pathways. Examples are provided of how the framework can be put into practice.

1 Introduction

Research and theory developed in information science (IS) has inspired many of the algorithms, models and interfaces developed and implemented in modern information retrieval (IR) systems [78]. Even with this influential link between the communities, quite recent commentary [17] suggests there is a gap in research between the two areas and a need for a holistic view of the searcher (the IS focus) and the search system (the IR focus) as one system together [36, 37]. In fact, Ingwersen and Järvelin [37] suggest this view must go beyond just the system and the searcher. Data science, a profession deemed "the sexiest job" [19], also has agency in the modern Web search environment. The data science community has quite an influential role in modern Web search environments, as members of this community are often tasked with development of models that are optimized to maximize user satisfaction [78], revenue and profits [19, 88] and user engagement [19]. Finally, there is the community of searchers that

seek and find information on the Web through a multitude of IR environments such as search engines (e.g. Google), product websites (e.g. Amazon) and social media news feeds (e.g. Facebook) that have their own biases and beliefs [7, 77]. Unfortunately, the biases and beliefs of the searcher and the IR systems they interact with form a feedback loop that not only changes the system [7, 8, 73] but also changes the beliefs of the user [7, 73, 77]. Ultimately, this self-reinforcing cycle exposes individual searchers and broader society to potentially harmful and dangerous outcomes [7, 55, 73, 77]. It is our view that the potential and already-realized harms caused by this reinforcing cycle are a side effect of the non-holistic view, a view which must extend beyond the searcher and the system [37], and thus motivates our framework for harm prevention in Web search.

Our framework is aimed at communities which include researchers in IS, IR, data science and the behavioral and cognitive sciences as well as the policy makers in governments and the leadership teams of Web platforms. There is common recognition by these communities of the ethical concerns and potentially grave implications of the technology that are now ubiquitous in our everyday lives. Simultaneously, aside from efforts by IS and IR [11, 17, 37], these communities appear to be working independently of one another. As such, we see a need for a common framework for all of these communities to jointly work towards a common goal of ethical responsibility to the searcher and broader society for which we are a part of. The components of the framework (Section 3) are ones we believe are the most essential for these communities to place focus initially. Components include policy updates, cognitive interventions, evaluation methods, and considerations for overall search system design. Central to our framework are four themes: collaborative effort by the communities mentioned, greater transparency to the user, greater choice for the user and an ethics-based approach for search system development.

2 Background

Why develop a framework? A number of researchers in the IR, IS and data science communities have expressed ethical concerns and potential for harmful ramifications due to the information that is collected and provided by the current Web search environments we deploy [7, 22, 35, 67, 77, 78]. In parallel, some researchers in our community have proposed and demonstrated methods that address some of these matters (see [10, 23, 27, 34, 64, 67, 87]). The behavioral and cognitive sciences community has voiced similar concerns [41, 43, 45] and offered possible solutions [41–43, 45] in line with the IR and IS communities. Yet, even though leadership of many popular Web platforms (where search for information commonly occurs) publicly express their concerns about these same matters, there are very few instances where they set policies for their platforms to align

with recommendations of researchers and policy makers[4] and the more common response is to take no action[5] until required by law to do so[6].

What are the harms? A broad spectrum of harms have occurred or have the potential to occur. Dangerous and potentially deadly health outcomes [77, 55], as well as destabilized political systems [14] have occurred. Excessive interaction (e.g. internet addiction) with information in social media environments is shown to have small but negative impacts to adolescents [53][7]. State-sponsored surveillance that monitors searcher behavior is sometimes used to harm individuals [13]. Other individuals interacting with IR systems have the potential of being radicalized and motivated to join extremist communities known to cause harm to others [73]. Users can receive dangerous advertisements based on their beliefs [62]. Indeed the harms can become quite dystopian [74] and these serve as the primary motivation for our proposal that aims to prevent individual and societal harms due to Web search.

What are the causes? The harms have many underlying causes, but we emphasize key factors. First, corporate (IR) platform policy can encourage data scientists to create addictive systems [19] and environments that are non-transparent to searchers[8] [41, 45]. There are of course the previously mentioned biases existent in both the system [7, 77] and the user [38, 51, 77], and the reinforcement factor [7, 8, 73]. Information itself is a factor. This comes in two forms, the information found (e.g. search results, Web pages, videos, social media post) and the information collected about the searcher (e.g. queries, IP address, usernames) – both of which offer many benefits to searchers such as exposure to more relevant information and faster task completion [78], but simultaneously may contain harmful content [77] and cost them their privacy [39, 78]. The centralized nature of search environments [82, 83], which are now commonplace, were built upon IS models of search developed and tested in quite different environments, such as the library [78, 83], and is another factor likely playing into concerns around privacy. The motive of profit [88] certainly encourages searchers to view information they might not otherwise do [74]. Moreover, the metrics used to evaluate systems (a focus of data scientists and IR researchers [80]) are quite different to those suggested by IS researchers (e.g. [21]). This issue of the communities (e.g. IR, data science and cognitive science) operating independently, rather than collectively towards a shared goal, is also a possible cause that should not be overlooked. The field of research known as interactive information retrieval (IIR)

[4] Exceptions include: Twitter now includes credibility assessments of claims for some Tweets [16]; Facebook moderates news feeds for hate speech [15].

[5] For example, nearly a decade has passed since [34, 64] suggested approaches to enable searchers a pathway for better assessments credibility, but to our knowledge, no commercial search engine has implemented these methods.

[6] In 2018 both the General Data Protection Regulation (GDPR) (to better protect privacy) [24] and NetzDG (to reduce racism and extremism) [63] forced major platforms to update their policies and practices or face stiff penalties.

[7] This study argues for better evaluation measures, which we discuss in Section 3.4

[8] For example, to understand privacy impacts, searchers must read lengthy privacy policy statements, written in an obfuscated manner.

aims for a collective and interdisciplinary view of the search process [36, 37] and it is through the IIR lens that the framework was developed and now introduced.

3 Framework Components

A recent proposal by Smith and Young Rieh [67] provides important insights for the development of a framework for Web search systems designed to prevent harms to individuals and society. Their proposal suggests that many users have information literacy and critical thinking skills that are useful to reduce the risks of harms from Web search. They highlight that current implementation methods of search systems, such as the Search Engine Results Page (SERP), offer little if any support to utilize these skills. Furthermore, they point out that users (likely due to the high cognitive demands of applying information literacy skills) put too much trust in the results found in the SERP, as has been demonstrated by other research [38, 77]. It appears some vitally important processes of search introduced by the IS community are being inhibited by the current design [67], processes including sense-making and exploratory search, which is unfortunate given the role they play in learning [78]. Their proposal also suggests that current IR systems are optimized for close-ended tasks (e.g. fact-based, question-answering), but should instead be optimized for learning. Ultimately, their proposal being that information in Web search interfaces should offer cues (e.g. topic, author and their affiliations, affective semantics including hate and humor) that enable users to utilize their literacy skills and assist them with critical thinking. However, their framework, as encompassing as it is, does not consider important matters such as privacy of the searcher [39] and platform policies optimized for corporate profit [80, 88]. That said, many elements of their proposal, such as considerations for interface design and informational cues that engage critical thinking for better decision making, are central to our proposal. We therefore see our work as an extension of their work. The decision-making factor is in fact fundamental to search [60, 78], suggesting a strong need for cognitive interventions, which are proposed as a bridge between many of the other framework components.

The proposed framework is segmented into four main areas which are seen as core to the development of search environments to reduce risk of harm to both the searcher and society. (1) Policy, which includes methods of law, education and corporate policy, are suggested. (2) A set of cognitive approaches, developed for the specific purpose of engaging decision making that reduces risk to individuals and society are introduced. (3) Considerations for the system design are provided, for which content enrichment and interface design are the main focus. (4) Any framework, and any approach for that matter, needs to be evaluated, for which suggestions are also given. The considerations provided are not exhaustive, and are intended as a foundation for a way forward.

3.1 Policy

Policy is a broad topic, which encompasses relevant areas such as law and education and can be used as a mechanism to prevent harm in Web search. Policies

set by the Web search systems are used as a means to leverage their commercial, legal and overall organizational interests. For instance, explicit privacy policies and data usage policies may be tailored to protect the provider from legal ramifications (e.g. the GDPR), while simultaneously maximizing their commercial profits [88]. Alternatively, a provider may shift their policy to meet social norms and address public outcry from issues such as misinformation [16]. There are also policy decisions around design choices for the core product that searchers interact with, such as how to present information in a SERP, search support tools to include in the product (e.g. query suggestion), and underlying retrieval and ranking models to implement. Clearly, web search platforms should keep harm prevention central to their design policy.

Laws can be implemented, locally, nationally, within economic regions and globally, with examples including California, Canada, European Union and human rights law respectively. For Web searchers, the GDPR [24], a law which is designed to better protect their privacy and allow for greater transparency and control over how data collected about them is used, is perhaps the most well known law for harm prevention to date. Laws may also be used to enforce information providers (e.g. social media and search platforms) to take down and / or filter out content that is perceived by law makers to be harmful to individuals and or societies. Examples include the NetzDG Germany [63] that require removal of speech that is hateful (e.g. Nazi imagery) and censoring of Google search results (e.g. websites mentioning the Tiananmen Square massacre) by the Chinese government [72]. Some laws, such as the communications and decency act in the USA [75], place the onus of legal liability on the publisher of the content (e.g. author of news article), but not the provider (e.g. search engine) or user (e.g. searcher). Legal tools achieve harm prevention through penalty (e.g. fines, imprisonment) for non-adherence to the rules stated by law, they also have a sense of authoritarianism and dictating what is good or bad. As differentiating between good and bad can be problematic [27], we suggest that law be used as a tool of last resort. Nonetheless, ethical considerations are a critical factor to IIR systems and research and therefore must be taken into account [40, 78]. As part of the framework, basic universal human rights [69–71] are the recommended lens through which policy is set.

Finally, education approaches and campaigns are a suggested pathway to improve search capabilities that minimize personal harm. There are some efforts by platforms to provide education tools and programs in primary and secondary schooling (see [30]) as well as being broadcast to searchers of any age (see [30, 59]). However, a searcher is not provided such tools directly in the search engines (i.e. there is no link provided)[9]. In our view, education is a promising pathway, as it overlaps with the cognitive interventions discussed in the section that follows.

[9] Query recommendations and spelling corrections may be educational if it can be shown that the user improves their query behavior or their spelling over time.

3.2 Behavioral and Cognitive Interventions

Decision making is fundamental to the search process [60, 78]. Therefore it is worthwhile to consider strategies from behavioral and cognitive sciences that are developed specifically for minimizing risk and harms. Three approaches, nudging [56, 68], boosting [33, 41] and techno-cognition [43], were recently proposed pathways to minimize and address harms in the modern online world [41, 45]. We focus on the first two approaches, nudging and boosting, as they are quite different in their methodology, yet very similar in their aim of reducing individual and societal risks. Additionally, we introduce nutrition labels and fact boxes as means to communicate potential harms.

Nudging [68] is a popular behavioral-public-policy approach, which has gained notoriety in recent years. Nudges aim to push people towards—what the 'nudger' believes to be—more beneficial decisions through the 'choice architecture' of people's environment (e.g., default settings). Thaler and Sunstein [68] provide the following definition of a *nudge*:

> A nudge ... is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not.

Nudging requires that the 'choice architecture' includes an element of libertarian paternalism [68], that is, the *nudge* must allow the individual to opt out (e.g., choose the non-default option); this is different from a purely paternalistic approach such as bans, which have no opt-out mechanism by design and intent. Some nudges, such as nutrition labels and warning lights, have educational elements [41, 86], but for the most part, nudges aim to directly change behavior without targeting people's competences [33]. The political philosophy and claims about human nature underlying nudging have been criticized recently [52, 28]; see, for example, [56] for a review of the issues discussed. Self-nudging [56] – people acting as their own "citizen choice architects" – has been proposed as a way to harness the power of nudging while largely circumventing its problems.

Boosting [33] is another approach to behavior change based on evidence from behavioral science. Quoting Hertwig and Grüne-Yanoff (p. 974):

> The objective of boosts is to improve people's competence to make their own choices. The focus of boosting is on interventions that make it easier for people to exercise their own agency by fostering existing competences or instilling new ones. Examples include the ability to understand statistical health information, the ability to make financial decisions on the basis of simple accounting rules, and the strategic use of automatic processes (...)

In the context of web search, *boosting* aims to improve people's skills to effectively and safely search the web. To achieve this, a boosting approach combines both IR research on how people search and adapt their search strategies to the environment [34, 42, 66] with general insights on human judgment decision making online [41, 45] and offline [29] to design and evaluate boosting interventions.

The key difference between a boosting and nudging approach lies in the former's assumption that people are not simply "irrational" (and thus need to be nudged towards better decisions), but that the human cognitive architecture is malleable and thus new competencies and skills can be instilled [33]—often requiring little time and effort. However, whether nudging or boosting is the "better" approach for a particular situation depends both on ethical considerations (e.g., how much value is placed on people's autonomy), but also on pragmatic considerations of which approach will likely be more successful in terms of effectiveness and economic and non-economic costs. For example, since boosting needs people's cooperation to be effective, boosting has the advantage that it—by design—cannot be manipulative. But this cooperation requirement also implies that boosting will not be successful in situations were people are unwilling or unable to learn or make use of a boost. See [32] for a discussion and some rules of thumb for when nudging or boosting is likely to work better.

Nutrition Labels and Fact Boxes Cognitive science has also provided a large body of evidence on visual approaches for communication of risk in an understandable way. Nutrition labels are one popular visual approach for risk communication, and it is shown that traffic light type approaches produce better outcomes and are more preferred by users [48, 65, 76] than tabular based approaches[10]. Originally designed for medical decision making for doctors and patients, fact boxes are another promising means to provide information in a manner that includes the benefits and harms of the available decisions [49][11]. Interestingly, both nutrition labels and fact boxes can act as the medium to perform a *nudge* or a *boost*.

3.3 Search System Design

Many components are necessary to build a fully functional search engine [18] and it is clear that the underlying systems have a tendency to become biased and steer users towards harmful information [7]. Here, we focus on content enrichment and the search interface and provide limited discussion on other components, such as log analysis and retrieval models hinting at where they might play a role within the framework. Based on commonly used implementation methods, such as those leveraging query logs as a primary means to model searchers and provide

[10] Such as those produced by the Food and Drug Administration https://www.fda.gov/food/nutrition-education-resources-materials/new-nutrition-facts-label

[11] See examples of fact boxes at the Harding Institute for Risk Literacy https://hardingcenter.de/en/fact-boxes

support tools (e.g. query recommendations, collaborative search models) [18, 78], it is conceivable that many of the system biases currently present [7] would abate over time due to the logging of interactions of a subset of users that make use of system elements discussed in sections below. That is, users that take the effort to minimize personal harm could in fact provide benefit to all users.

Content Enrichment with Informational Cues Processes that enrich and classify information in Web documents are fundamental to modern search engines and IR systems [18]. As previous research indicates, there are many different cues to consider [27, 42, 67] during this enrichment process to be applied to information that may be useful for minimizing risk to searchers, for which many are important factors for making better decisions in search [34, 42, 66].

For IR researchers and data scientists, the listing of cues provided by Smith and Rieh (see [67]) as well as methods outlined by Fuhr et al (see [27]) are useful guides for development of cue extraction methods. Methods are already available for extracting cues such as the reading level, the virality of the content (i.e. how likely will the information spread), emotionality (e.g. language that is angry, overly positive, etc.), prevalence of factual, opinionated and / or controversial information, trustworthiness of the source (e.g. mechanisms to determine the credibility of a Web page), technicality (e.g. a score for amount of technical jargon in document) and if the document is currently topically relevant [27].

Bibliographic cues (e.g. author affiliation) and inferential cues (e.g. citations to and from document) are also needed for critical thinking and evaluation of information [67]. A lack of transparency exists in affiliations of authors and publishers of information [67], and therefore there is a clear need for developing methods that evaluate affiliation(s) of authors and publishers of information (e.g. who is funding think tank X that publishes web page Y) [43].

Methods to identify content that is hateful [84], misogynistic [9] or containing vulgar language [20] are also readily available and potentially useful for minimizing exposure to content that some users may find offensive, which Smith and Rieh [67] classify as valence cues. Marking content which is sexually explicit (written, verbally and / or visually) [54], may be useful for developing strategies to minimize harms to minors as well as users that are susceptible to addiction.

As privacy is of paramount concern too, extracting cues that provide greater transparency to the searcher into what data is collected and by whom it is collected and shared are also critical to prevent harms from the collected information. One such task in this space is the identification of 3rd parties that data will be shared with when visiting a Web page [44, 81] and another being the classification of privacy statements on the websites where the content is hosted, a task that could be designed with existing privacy statement corpora [79]. In a similar vein as author affiliation cues proposed by [67] and [43], privacy-based ontologies containing information (e.g. total fines, number of GDPR violations) about 1st-party providers and the 3rd-party affiliations could be developed to present privacy cues.

Many of the cues can be extracted with models produced by machine learning algorithms. Nevertheless, for data scientists that develop these models, it is critical to minimize model bias, such as gender bias [12]. Many solutions in the field of IR are designed with a "find the best" model mindset, as evidenced by the leaderboard approach for shared tasks (e.g. TREC, SemEval), and are a likely cause of some model biases and subsequent poor predictions. There is evidence that ensemble approaches are more robust and resilient to bias and more likely to outperform a single model [31, 46, 84], and are one possible alternative. In search spaces with potentially dangerous outcomes (e.g. health), data scientists should also consider interpretable models [58].

Interface Design Informational cues, cognitive interventions and policy are all important for harm reductions [45], but they need a medium for implementation and it is the search interface (such as a SERP) that is this medium.

Extracting cues that allow for the design of better decision making tools (thus enabling users to better tap into their critical thinking skills) and designing interfaces that present such cues and tools in a not-too-disruptive manner are two major challenges for interface design. Commercial SERPs are typically presented as ranked lists [67] and, depending on the query, will contain content such as advertisements, social media posts and news articles [5]. Search support tools are an important IR system component for improving search [78], some of which are available within the SERP including query suggestions and auto-completion as well as spelling correction. Thus, any component that allows the user to minimize the chance of harm, also falls within the scope of search support.

Space is a premium and one challenge is to ensure that the screen is not overloaded [47]. Risk communication tools such as nutrition labels and fact boxes are highly effective and desirable, but may not fit on small mobile devices, where warning lights are likely the better option. Link enrichment is another approach [47], where pop-ups populated with informational cues are included with the results, and is thus especially appealing as it could be applied to both desktop and mobile search. Link enrichment also need not apply only to the SERP, and can be applied as searchers navigate within [2] or across domains and the Web (Wikipedia desktop offers link enrichment and is one live example).

Alternatively, the SERP could be designed to rank or filter results as to attenuate possible harms from, say, privacy concerns or dangerous medical advice [87]. Indeed, commercial search engines already offer the default of filtering adult content (e.g. content that is classified as sexually explicit), and takes up little space within the interface. Altering results in the SERP in this manner is a *nudge*, so long as the user is given the capability to opt-out [68]. However, we caution against such approaches, as it does not tap into the important critical thinking and literacy skills of searchers [67] and thus likely does not generalize to other contexts without such a nudge.

The interface is also where policy can be implemented. It is conceivable that law makers may someday require IR systems to include any number of the approaches already discussed—an information nutrition label [27] is one such pos-

sibility. Or platforms, such as Google, could voluntarily set policy that provides links to educational resources in the SERP, simplifying the process for searchers to learn how to better protect themselves during the search process.

3.4 Evaluation

Evaluation is a fundamental and necessary process for IIR research. There are many resources readily available (for an introduction see [40] for evaluation of user studies and [18] for evaluation of IR systems) to perform evaluations of IIR systems and user interactions. Traditional evaluation metrics, such as relevance-based metrics (e.g. precision, F-measure), and human-based metrics (e.g. time to complete tasks, query abandonment rate) are essential for harm prevention strategies developed with this framework, as there should be minimal impact on these metrics. For data scientists and other researchers that perform evaluations, they should consider recent suggestions of leading experts (see [26, 61]), as IR studies often lack statistical rigor [61] and that many easily avoidable mistakes are made [26]. We now turn our attention to more recently proposed metrics that should also be considered, including metrics that take an economic view, and a somewhat newer generation of outcome-based metrics.

Economic-Based Metrics Interventions that reduce risk of harm, such as those suggested in the framework, have costs (e.g. time) for the individual [32, 68] and costs are an important economic consideration for IIR environments [4, 6][12]. The economic view has inspired a new set of useful evaluation approaches, which integrate theories from economics and have the overall aim to better predict user behavior in the search environment [3]. Incorporation of the economic view of IIR is potentially useful for evaluating the framework, as it allows evaluation from the perspective of trade-offs of costs and benefits [3, 5], such as the trade-off of costs of time for the benefit of reduced risk of harm as part of the search process. In addition to time, examples of relevant costs one might consider are the money a searcher is willing to pay for information that is of high quality, amount of data they are willing to share with 3rd parties and the effort of searching for information relevant to their task.

Outcome-Oriented Metrics Sense-making, one area of research within IS that considers the process of filling in gaps of knowledge, also has a strong focus on the ultimate outcome of this process [21], outcomes which may have positive or negative impacts [21, 57]. Such impacts fall in the domain of success metrics [78], which are possibly the most important evaluation approach for the framework, as they can be measured from the user perspective (does the user believe their risk of harm was reduced) and from the system perspective (did user X, making use of a privacy intervention, share less data than user Y, who

[12] There are costs with respect to designing and operationalizing interventions in a search environment, such as salaries for software engineers. However, for this discussion, we are strictly concerned with the economics of the searcher.

did not). Such metrics are important in the area of health search as incorrect information (resulting in incorrect knowledge) risks great harm [55, 77]. Also promising are new evaluation methods, such as overall reputation of commercial search platforms and longitudinal studies that sample regular users[13].

4 The Framework in Practice

Several empirical studies and commercial systems have some (but not all) of the elements of the framework. It is worth noting these to provide a lens into how the framework can be used in practice. To our knowledge, however, no approach addresses all four components. *Behavioral interventions* (*nudges*) and system-based *content enrichment* were shown to effectively steer users towards healthier food choices [23] and away from Websites that more greatly impact personal privacy [1, 87]. *Outcome-oriented evaluation* measures were considered in the latter study [87], but lack the *policy* element. Some commercial search engines (e.g. DuckDuckGo), have used *policy*, *system design* and *cognitive* approaches to protect users from adult material, but do not publish *evaluation* approaches.

Specific to *behavioral and cognitive interventions*, there are additional empirical findings worth noting. A subset of the cues suggested by Smith and Rieh [67] were used to augment search results visually to *nudge* users to more accurately assess credible information [64]. Browser plug-ins can provide a visual *nudge* during Web browsing and exploration, such as the Ghostery 3rd-party blocking tool (https://www.ghostery.com/), which by default blocks data sharing with 3rd parties[14]. In line with a *boosting* approach, one study tested low-cost search tips as a means to provide skills for better searching [50] and another study improved novice searchers skills by feedback based their search behavior compared to expert searchers [10]; note that neither study was explicitly designed for harm reduction nor explicitly referred to *boosting*.

Evidence suggests that elements in the URL can be utilized to *boost* users with a skill to better protect privacy and simultaneously improve health outcomes [85] and Figure 1 is a prototype of a *boost* that combines these findings with a fact box [49] as a means to reduce risks of 3rd-party tracking. Commercial search platforms may someday shift their policy to offer more focus on harm prevention, where policy might place such a fact box in the SERP, or alternatively offer tools allowing users to *self-nudge* [56] (e.g. selecting search domains, such as health and politics, to filter out non-credible information by default).

5 Conclusions

We introduced a framework as a pathway to reduce the risk of harms present in modern Web search. Central to the framework are cognitive decision making tools and three further components: policy, system design and overall evaluation.

[13] The Harvard nurse study https://www.nurseshealthstudy.org/ may be a useful template for longitudinally assessing harm and risk factors of search systems.

[14] Caution should prevail with 3rd-party blocking tools as recent findings question their effectiveness [25].

Out of 100 websites visited for health & medical issues in a search engine:		
How many out of 100:	**.gov/.org websites**	**other websites**
Benefits		
will share your data with ≤ 2 3rd party companies?	59	14
Harms		
will share your data with ≥ 8 3rd party companies?	11	33

Fig. 1. A proposed *boost* based upon fact box methods [49] and previous findings on the usefulness of the top-level domain of a URL for harm reduction [85]. Data scientists and IR researchers could make use of sampling techniques to dynamically populate fact boxes based upon the topical search space of a user's information need.

Baeza-Yates' recent commentary on the interactions between IR systems and searchers [7] as a cause of harms in Web search systems, is direct real-world evidence of the pervasive nature of the current Web search setup. Implementing the strategies sketched in this article is a possible approach to improving the overall system. Given current algorithms and their ability to learn from log data, hypothetically it would only require a subset of users concerned about harms in Web search to shift the results for everyone else (i.e., a positive externality). The environment could naturally evolve to something more protective for individuals and society as a whole.

We recognize that our framework is just a start, and there is much more to be considered, especially in the space of ensuring overall commercial value and limiting the impacts on system performance. Embedding the current, initial framework in the IIR view, which takes a multi-faceted and interdisciplinary approach inclusive of these additional factors, will ensure that it can mature into a comprehensive and realistic framework. A key message from our proposal is that methods from cognitive science introduced in our framework appear particularly promising [33, 41, 45]. We hope that such a framework will be a useful discussion point for members of all communities involved to work towards a common goal of minimizing the harm for searchers and our society.

Acknowledgements Thank you to the reviewers for their constructive feedback that has helped strengthen this paper. This work was supported by the Economic and Social Research Council grant number ES/M010236/1 (The Human Rights Big Data and Technology (HRBDT) Project).

References

1. Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L.F., Komanduri, S., Leon, P.G., Sadeh, N., Schaub, F., Sleeper, M., Wang, Y., Wilson, S.: Nudges for privacy and security: Understanding and assisting users' choices online. ACM Comput. Surv. **50**(3) (August 2017)
2. Alhindi, A., Kruschwitz, U., Fox, C., Albakour, M.D.: Profile-based summarisation for web site navigation. ACM Transactions on Information Systems (TOIS) **33**(1), 4:1–4:39 (Mar 2015)
3. Azzopardi, L.: The economics in interactive information retrieval. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 15–24. SIGIR '11 (2011)
4. Azzopardi, L.: Modelling interaction with economic models of search. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–12. SIGIR '14, ACM (2014)
5. Azzopardi, L., Thomas, P., Craswell, N.: Measuring the utility of search engine result pages: An information foraging based measure. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 605–614. SIGIR '18 (2018)
6. Azzopardi, L., Zuccon, G.: An analysis of the cost and benefit of search interactions. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. pp. 59–68. ICTIR '16 (2016)
7. Baeza-Yates, R.: Bias on the web. Communications of the ACM **61**(6), 54–61 (2018)
8. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on facebook. Science **348**(6239), 1130–1132 (2015)
9. Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M., et al.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics (2019)
10. Bateman, S., Teevan, J., White, R.W.: The search dashboard: How reflection and comparison impact search behavior. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 1785–1794. CHI '12, Association for Computing Machinery (2012)
11. Belkin, N.J.: Some(what) grand challenges for information retrieval. SIGIR Forum **42**(1), 47–54 (2008)
12. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: Advances in Neural Information Processing Systems 29, pp. 4349–4357. NIPS, Curran Associates, Inc. (2016)
13. Botsman, R.: Big data meets big brother as china moves to rate its citizens. Wired UK **21** (2017)
14. Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian **17**, 22 (2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
15. Carlsen, A., Haque, F.: What Does Facebook Consider Hate Speech? Take Our Quiz. https://www.nytimes.com/interactive/2017/10/13/technology/facebook-hate-speech-quiz.html (October 2017), (Accessed on 10/21/2018)

16. Conger, K., Alba, D.: Twitter Refutes Inaccuracies in Trump's Tweets About Mail-In Voting - The New York Times. https://www.nytimes.com/2020/05/26/technology/twitter-trump-mail-in-ballots.html (May 2020), (Accessed on 06/14/2020)
17. Croft, W.B.: The importance of interaction for information retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1–2. SIGIR'19, ACM, New York, NY, USA (2019)
18. Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Pearson Education (2015)
19. Davenport, T.H., Patil, D.: Data scientist: The sexiest job of the 21st century. Harvard business review **90**(5), 70–76 (2012)
20. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009 (2017)
21. Dervin, B.: Sense-making theory and practice: an overview of user interests in knowledge seeking and use. Journal of Knowledge Management **2**(2), 36–46 (1998)
22. Diaz, F.: Worst practices for designing production information access systems. ACM SIGIR Forum **50**(1), 2–11 (2016)
23. Elsweiler, D., Trattner, C., Harvey, M.: Exploiting food choice biases for healthier recipe recommendation. In: Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 575–584. SIGIR '18, ACM (2017)
24. European Union: General Data Protection Regulation (GDPR) – Official Legal Text (2016 April), https://gdpr-info.eu/, (Accessed on 04/02/2018)
25. Fouad, I., Bielova, N., Legout, A., Sarafijanovic-Djukic, N.: Missed by filter lists: Detecting unknown third-party trackers with invisible pixels. Proceedings on Privacy Enhancing Technologies **2020**(2), 499–518 (2020)
26. Fuhr, N.: Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. ACM SIGIR Forum **51**(3), 32–41 (2017)
27. Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdl, W., Peters, I., Stein, B.: An information nutritional label for online documents. ACM SIGIR Forum **51**(3), 46–66 (2017)
28. Gigerenzer, G.: On the supposed evidence for libertarian paternalism. Review of philosophy and psychology **6**(3), 361–383 (2015)
29. Gigerenzer, G., Hertwig, R., Pachur (Eds.), T.: Heuristics: The Foundations of Adaptive Behavior. Oxford University Press (2011)
30. Google: Search education – google. https://www.google.com/insidesearch/searcheducation/, (Accessed on 06/14/2020)
31. Hagen, M., Potthast, M., Büchner, M., Stein, B.: Webis: An ensemble for twitter sentiment detection. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 582–589 (2015)
32. Hertwig, R.: When to consider boosting: some rules for policy-makers. Behavioural Public Policy **1**(2), 143–161 (2017)
33. Hertwig, R., Grüne-Yanoff, T.: Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science **12**(6), 973–986 (2017)
34. Hilligoss, B., Rieh, S.Y.: Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing & Management **44**(4), 1467 – 1484 (2008)
35. Horvitz, E., Mulligan, D.: Data, privacy, and the greater good. Science **349**(6245), 253–255 (2015)

36. Ingwersen, P.: Cognitive perspectives of information retrieval interaction: Elements of a cognitive ir theory. Journal of Documentation **52**(1), 3–50 (1996)
37. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context, vol. 18. Springer Science & Business Media (2005)
38. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 154–161. SIGIR '05, ACM (2005)
39. Jones, K.S.: Privacy: what's different now? Interdisciplinary Science Reviews **28**(4), 287–292 (2003)
40. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Foundations and Trends® in Information Retrieval **3**(1–2), 1–224 (2009)
41. Kozyreva, A., Lewandowsky, S., Hertwig, R.: Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest (In Press 2020)
42. Lee, M.D., Loughlin, N., Lundberg, I.B.: Applying one reason decision-making: the prioritisation of literature searches. Australian Journal of Psychology **54**(3), 137–143 (2002)
43. Lewandowsky, S., Ecker, U.K., Cook, J.: Beyond misinformation: Understanding and coping with the "post-truth" era. Journal of Applied Research in Memory and Cognition **6**(4), 353 – 369 (2017)
44. Libert, T.: Privacy implications of health information seeking on the web. Communications of the ACM **58**(3), 68–77 (2015)
45. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C.R., Hertwig, R.: How behavioural sciences can promote truth, autonomy and democratic discourse online. Nature Human Behaviour (2020)
46. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PloS one **14**(8) (2019)
47. Marchionini, G., Geisler, G., Brunk, B.: Agileviews: A human-centered framework for interfaces to information spaces. In: Proceedings of the Annual Conference of the American Society for Information Science. pp. 271–280 (2000)
48. Maubach, N., Hoek, J., Mather, D.: Interpretive front-of-pack nutrition labels. comparing competing recommendations. Appetite **82**, 67–77 (2014)
49. McDowell, M., Rebitschek, F.G., Gigerenzer, G., Wegwarth, O.: A simple tool for communicating the benefits and harms of health interventions: A guide for creating a fact box. MDM Policy & Practice **1**(1) (2016)
50. Moraveji, N., Russell, D., Bien, J., Mease, D.: Measuring improvement in user search performance resulting from optimal search tips. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 355–364. SIGIR '11, Association for Computing Machinery (2011)
51. Novin, A., Meyers, E.: Making sense of conflicting science information: Exploring bias in the search engine result page. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 175–184. CHIIR '17, ACM (2017)
52. Oliver, A.: Nudging, shoving, and budging: Behavioural economic-informed policy. Public Administration **93**(3), 700–714 (2015)
53. Orben, A.: Teenagers, screens and social media: a narrative review of reviews and key studies. Social Psychiatry and Psychiatric Epidemiology pp. 1–8 (2020)
54. Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., Rocha, A.: Video pornography detection through deep learning techniques and motion information. Neurocomputing **230**, 279 – 293 (2017)

55. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.: The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. pp. 209–216. ICTIR '17, ACM (2017)
56. Reijula, S., Hertwig, R.: Self-nudging and the citizen choice architect. Behavioural Public Policy p. 1–31 (2020). https://doi.org/10.1017/bpp.2020.5
57. Reinhard, C.D., Dervin, B.: Comparing situated sense-making processes in virtual worlds: Application of dervin's sense-making methodology to media reception situations. Convergence **18**(1), 27–48 (2012)
58. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
59. Russell, D., Callegaro, M.: How to be a better web searcher: Secrets from google scientists researchers who study how we use search engines share common mistakes, misperceptions and advice. Scientific American (2019), https://blogs.scientificamerican.com/observations/how-to-be-a-better-web-searcher-secrets-from-google-scientists/
60. Ruthven, I.: Interactive information retrieval. Annual Review of Information Science and Technology **42**(1), 43–91 (2008)
61. Sakai, T.: Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 5–14. SIGIR '16, ACM, New York, NY, USA (2016)
62. Sankin, A.: Want to Find a Misinformed Public? Facebook's Already Done It. https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it (April 2020), (Accessed on 04/30/2020)
63. Schulz, W.: Regulating Intermediaries to Protect Privacy Online – the Case of the German NetzDG. Personality and Data Protection Rights on the Internet (2018)
64. Schwarz, J., Morris, M.: Augmenting web pages and search results to support credibility assessment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 1245–1254. CHI '11, Association for Computing Machinery (2011)
65. Siegrist, M., Leins-Hess, R., Keller, C.: Which front-of-pack nutrition label is the most efficient one? the results of an eye-tracker study. Food Quality and Preference **39**, 183–190 (2015)
66. Smith, C.L., Kantor, P.B.: User adaptation: Good results from poor systems. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 147–154. SIGIR '08, Association for Computing Machinery, New York, NY, USA (2008)
67. Smith, C.L., Rieh, S.Y.: Knowledge-context in search systems: Toward information-literate actions. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. pp. 55–62. CHIIR '19, ACM, New York, NY, USA (2019)
68. Thaler, R.H., Sunstein, C.R.: Nudge: Improving Decisions about Health, Wealth, and Happiness. Penguin (2009)
69. The United Nations: Universal declaration of human rights (1948)
70. The United Nations General Assembly: International covenant on civil and political rights. Treaty Series **999**, 171 (December 1966)
71. The United Nations General Assembly: International covenant on economic, social, and cultural rights. Treaty Series **999**, 171 (December 1966)
72. Thompson, C.: Google's China Problem (and China's Google Problem). The New York Times (2006), (Accessed on 06/2020)

73. Tufekci, Z.: Facebook said its algorithms do help form echo chambers, and the tech press missed it. New Perspectives Quarterly **32**(3), 9–12 (2015)
74. Tufekci, Z.: We're building a dystopia just to make people click on ads (September 2017), https://www.ted.com/
75. U.S. Government Publishing Office: 47 usc 230: Protection for private blocking and screening of offensive material. https://www.govinfo.gov/content/pkg/USCODE-2011-title47/pdf/USCODE-2011-title47-chap5-subchapII-partI-sec230.pdf (1996), (Accessed on 06/2020)
76. Van Herpen, E., Van Trijp, H.C.: Front-of-pack nutrition labels. their effect on attention and choices when consumers have varying goals and time constraints. Appetite **57**(1), 148–160 (2011)
77. White, R.: Beliefs and biases in web search. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–12. SIGIR '13, ACM (2013)
78. White, R.W.: Interactions with Search Systems. Cambridge University Press (2016)
79. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 1330–1340 (2016)
80. Young, K.S.: Internet addiction: A new clinical phenomenon and its consequences. American Behavioral Ccientist **48**(4), 402–415 (2004)
81. Yu, Z., Macbeth, S., Modi, K., Pujol, J.M.: Tracking the trackers. In: Proceedings of the 25th International Conference on World Wide Web. pp. 121–132. WWW '16, International World Wide Web Conferences Steering Committee (2016)
82. Zimmer, M.: Privacy on Planet Google: Using the Theory of Contextual Integrity to Clarify the Privacy Threats of Google's Quest for the Perfect Search Engine Google: An Intersection of Business and Technology. Journal of Business & Technology Law **3**, 109–126 (2008)
83. Zimmer, M.: Web Search Studies: Multidisciplinary Perspectives on Web Search Engines, pp. 507–521. Springer Netherlands (2010)
84. Zimmerman, S., Fox, C., Kruschwitz, U.: Improving hate speech detection with deep learning ensembles. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (Miyazaki, Japan). LREC 2018 (2018)
85. Zimmerman, S., Thorpe, A., Chamberlain, J., Kruschwitz, U.: Towards search strategies for better privacy and information. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. pp. 124–134. CHIIR '20, Association for Computing Machinery (2020)
86. Zimmerman, S., Thorpe, A., Fox, C., Kruschwitz, U.: Investigating the interplay between searchers' privacy concerns and their search behavior. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 953–956. SIGIR'19, ACM (2019)
87. Zimmerman, S., Thorpe, A., Fox, C., Kruschwitz, U.: Privacy nudging in search: Investigating potential impacts. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. pp. 283–287. CHIIR '19 (2019)
88. Zuboff, S.: Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology **30**(1), 75–89 (Mar 2015)