FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval

Editors
Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, and Michael D. Ekstrand

Authors and participants
Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W. Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N. Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D. Ekstrand, Preethi Lahoti, and Toshihiro Kamishima

Abstract

The purpose of the SIGIR 2019 workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety (FACTS-IR) was to explore challenges in responsible information retrieval system development and deployment. To this end, the workshop aimed to crowdsource from the larger SIGIR community and draft an actionable research agenda on five key dimensions of responsible information retrieval: fairness, accountability, confidentiality, transparency, and safety. Such an agenda can guide others in the community that are interested in pursuing FACTS-IR research, as well as inform potential funders about relevant research avenues. The workshop brought together a diverse set of researchers and practitioners interested in contributing to the development of a technical research agenda for responsible information retrieval.

1 Introduction

Information retrieval (IR) systems and related technologies, such as search engines, recommender systems, and conversational assistannts, are responsible for organizing, curating, and promoting most of the information that is being consumed today. Importantly, IR systems are not isolated systems: they reflect the content and interaction data used to develop them and their impact on the environments in which they operate. Indeed, IR systems connect people to information, shaping not only the information consumption patterns, but also the social interactions, affecting both what and who is more visible and when are they visible

(Biega et al., 2018; Stoica et al., 2018; Hannák et al., 2017; Nilizadeh et al., 2016; De-Arteaga et al., 2019).

Recognition of the social and political implications of information retrieval goes back at least two decades (Friedman and Nissenbaum, 1996; Introna and Nissenbaum, 2000). More recent empirical evidence shows, for instance, that there are gaps in access to information across communities, in part due to the information needs of certain communities being less supported than those of others—often the more dominant communities (Goldman, 2005; Spink and Zimmer, 2008; Mehrotra et al., 2017; Golebiewski and boyd, 2018; Stoica et al., 2018). As with other AI-driven technologies, IR systems are also under the influence of those that design, build, maintain or use them, embedding and amplifying their biases (Barocas and Selbst, 2016; Olteanu et al., 2019a; Baeza-Yates, 2018; Hutchinson and Mitchell, 2019). Failures of IR systems may not always be easily traceable (Sharchilev et al., 2018) and the extensive use of interaction logs may lead to undesirable leaking of sensitive, secrete information (Yang et al., 2016). While users are now entitled to explanations of algorithmic decisions in certain parts of the world (EU, 2016), it is unclear how explanations, evidence trails and provenance might be communicated to the various user groups, and how such communications might change behaviors, and the quality, quantity, and nature of user interaction with IR systems (Dietvorst et al., 2015; Manjoo, 2015; Miller et al., 2017). Resilience to manipulation by external parties is also increasingly critical across a growing number of application scenarios (Wang et al., 2018).

Such fundamental issues concern all aspects of IR system development and deployment. Given the current ubiquitous use of a variety of IR systems, from web search to recommendation platforms to personal assistants, they have potentially wide ranging impact—both positive and negative. We know that people are more likely to trust sources ranked higher in the search or recommendation results, but the recommendation or ranking criteria might optimize for the perceived user satisfaction, possibly at the expense of providing factual information (Schwarz and Morris, 2011; White, 2013). For consequential user tasks, such as those related to medical, educational, or financial outcomes, this raises concerns about potential harms, and what the right trade-offs might be.

Over the last years, a community has coalesced to address questions of fairness, accountability, transparency, ethics, and justice in machine learning and other computing systems. The FACTS-IR workshop aimed to give that discussion a home at SIGIR 2019 and provide an opportunity to highlight challenges specific to IR systems (Olteanu et al., 2019c).

1.1 Workshop Format

The FACTS-IR workshop aimed to both provide a venue for work-in-progress as well as identify gaps in the emerging technical work on responsible IR, including under-theorized and under-specified issues related to each of the five FACTS-IR focus areas, in order to create actionable technical research agendas for each of them. We supported these goals with a two-part program consisting of presentations (of both workshop submissions and invited talks) and breakout group discussions, as follows:

- Morning: The first part featured presentations, based on both the submissions accepted by our PC members, as well as invited talks on each of the 5 pillars of the workshop (described in Section 1.1.3).
- Afternoon: In the second part, the participants were organized in working groups and were tasked with articulating a research agenda aiming at identifying priorities for one

of the FACTS-IR topics.

Each part was organized around one of the 5 topical pillars, with the afternoon session collecting input for a community-driven research agenda on FACTS-IR topics, by answering the following questions for each of them: (1) Which topics? (topics selection and description) (2) Why does the topic matter? (3) How is it relevant for IR? (4) What are the main research questions around the topic? (5) What are the key challenges and obstacles? (6) How will this make an impact? On whom?

1.1.1 Paper & invited talks

We solicited submissions as both full (8-page) research papers and 2–4 page extended abstracts as position papers that address issues related to the five FACTS-IR topical pillars. The submissions we received primarily touched on two of the areas of interest (accountability and transparency) with a majority covering issues related to transparency in textual summarization, deriving global explanations through the aggregation of local explanations, understanding and explaining predictions from tree-based boosting ensembles, and making user bias explicit in fact checking tasks. Other submissions discussed efforts to understand and define fairness metrics in IR systems, including tasks like ranking and recommendations, as well as applications to judicial systems.

Further, to cover the remaining focus areas, the workshop featured additional short presentations by academic and industry practitioners who are leaders in the FACTS research areas. The paper proceedings are available at (Olteanu et al., 2019b), and listed below along with the invited talks:

Fairness

- (Talk) Preethi Lahoti (MPI-INF) – *Operationalizing Individual Fairness for Algorithmic Decision Making*
- (Paper) Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow (Google) – *Fairness in Recommendation Ranking through Pairwise Comparisons*

Accountability

- (Talk) Maria Panteli (BBC Datalab) – *Accountability and Recommendation Systems at BBC*
- (Paper) Anubrata Das, Kunjan Mehta, and Matthew Lease (University of Texas at Austin) – *CobWeb: A Research Prototype for Exploring User Bias in Political Fact-Checking*

Transparency

- (Talk) Krisztian Balog (University of Stavanger & Google) – *Questions around Transparency in Information Retrieval*
- (Paper) Joris Baan, Maartje ter Hoeve (University of Amsterdam), Marlies van der Wees, Anne Schuth (De Persgroep, Amsterdam), and Maarten de Rijke (University of Amsterdam) – *Do Transformer Attention Heads Provide Transparency in Abstractive Summarization?*
- (Paper) Ana Lucic (University of Amsterdam), Hinda Haned (Ahold Delhaize), and Maarten de Rijke (University of Amsterdam) – *Explaining Predictions from Tree-based Boosting Ensembles*
- (Paper) Ilse van der Linden (University of Amsterdam), Hinda Haned (Ahold Delhaize), and Evangelos Kanoulas (University of Amsterdam) – *Global Aggregations of Local Explanations for Black Box models*

Confidentiality

- (Talk) Mahmoud F. Sayed (University of Maryland) – *Search Among Sensitive Content*
- (Paper) Graham McDonald, Craig Macdonald, and Iadh Ounis (University of Glasgow) – *The FACTS of Technology-Assisted Sensitivity Review*

Safety & Ethics

- (Talk) Pierre-Nicolas Schwab (RTBF) – *AI, Ethics, and Information*

1.1.2 The How: Breakout groups

The afternoon segment of the workshop consisted of focused discussions to articulate new research agendas and identify open problems in the FACTS-IR space, which we organized into working groups. The goal of each working group was the formulation of in-depth, concrete research agendas for each of the five areas, as well as the identification of potential tensions between them.

Working groups consisted of 3–6 participants, with participants voting and volunteering for one of the proposed topics, on a first come-first served basis.

1.1.3 The What: Proposed topics

The FACTS-IR workshop covered five key areas of focus, building on the responsible IR agenda articulated in the SWIRL report (Allan et al., 2018):

- **Fair IR**: the IR system should avoid discrimination across people and communities. To do so the notion of fairness should be contextual and well-grounded in the application, problem, and domain. Achieving fairness may be further complicated by the multi-stakeholder nature of most IR systems.
- **Accountable IR**: the IR system should be able to justify its recommendations or actions to users and other stakeholders, as well as be reliable at all times. This requires an understanding of the potential harms of using the system and of who is more likely to be affected. It also requires recourse avenues and processes for redress.
- **Confidential IR**: the output or actions of the IR system should not reveal secrets. IR systems often combine extensive behavioral logs to model their users, which if not properly handled can result in unintended leakage of information.
- **Transparent IR**: the IR system should be able to explain to users and other interested stakeholders why and how the suggested results were obtained. Providing proper explanations may require answering who the users and the stakeholders are. More broadly, the IR systems should be able to enable third parties to monitor and probe that the systems behave as expected.
- **Safe IR**: The IR system should be resilient to manipulation by possible adversarial parties, and should not expose the users to undesirable, harmful content.

Before the workshop (and inspired by the paper presentations and invited talks), the organizers identified possible sub-topics for each of the pillar topics listed above, which the participants volunteered for. In addition, before asking the participants to select one of these sub-topics, we solicited feedback on the proposed sub-topics; one outcome of this exercise was to add an additional sub-topic under the **Confidential IR** pillar, dubbed *Data biases in sensitive information*. The final sub-topics are listed below:

Fair IR

(1) Auditing IR systems for fairness

(2) *Fair-IR metrics and definitions*
(3) *Build Fair-IR*
Accountable IR
(4) *Understanding harms and their impact*
(5) Build auditable IR
(6) Recourse avenues and processes for redress
Confidential IR
(7) *Information leakage scenarios*
(8) Build confidential-IR
(9) *Data biases in sensitive information*
Transparent IR
(10) *Build transparent-IR*
(11) *Transparency dimensions*
Safe IR
(12) Resilience to manipulation
(13) *Limit exposure to harmful content*
(14) Build safe-IR

2 Seven FACTS-IR Topics Worked Out

Out of a total 14 proposed sub-topics for the breakout groups, based on participants voting preferences, 7 sub-topics were selected for the afternoon working groups to focus on, including two from the Fair-IR pillar, two from the Confidential-IR pillar, two from the Transparent-IR, while the last break-out group topic combined two sub-topics from the Safe-IR and the Accountable-IR pillars focused on harms, namely *Harms in IR Systems*. The selected sub-topics are *emphasised* in the list above, and detailed in this section as separate sub-sections.

2.1 Fair-IR Metrics and Definitions

Although fairness is arguably the basis of democratic societies, there is no one-fits-all definition of fairness—this also holds for Fair-IR. It is difficult to guarantee fairness in IR systems without measuring it, and it is difficult to measure fairness without a clear definition. Fairness is also multi-dimensional (similar to relevance), being dependent on the domain, time (follows standards of morality), context, topic, and stakeholder (users, society, company). These lead to challenges in developing theories that can inform how tools that can identify and correct bias should be built, and in understanding the associated trade-offs between exposure and impact for each such tool. Maybe the most important obstacles are the general lack of standards and data, as well as the costs.

2.1.1 Why is it important?

As our world becomes increasingly present online, so do our biases, inequalities, and historical gaps through the information we upload and consume on the Internet. While information retrieval systems are crucial in providing rapid access to information, they unfortunately also encode and sometimes amplify such gaps (e.g., in the way we perceive gender, race, political content, and so on).

Fairness, as seen in recent Computer Science literature, does not have a one-size-fits-all definition, but it is context dependent: what is personalization in one case, may be unequal access to opportunities in another. For that reason, we strive to operationalize the concept of fairness depending on the application, to come up with definitions and tools that help us identify the issues at stake, and to build systems that are inherently fairer.

2.1.2 How is this relevant to IR?

Information retrieval deals with the storage, organization, and retrieval of information. As such, it can implicitly encode existing biases, such as representation bias based on what information is available as well as bias that may inadvertently result from the technical decisions regarding information storage. Moreover, ranking systems may be prone to amplifying existing biases (for example, position bias can contribute to a rich-get-richer phenomenon), while recommender systems can lead to self-fuelling filter bubbles, contributing to growing societal biases.

Since IR systems have such prolific roles in our modern society by governing access to a wide range of information, they can impact human decision making as well as influence the health of our society. Thus, fairness in IR systems becomes of utmost importance for ensuring equal treatment and chances for individual and groups to access information and to gain exposure.

As IR systems by design aggregate massive datasets, measurements over such datasets provide us with the opportunity to discover large-scale inequality and also mitigate discrimination through re-designing our systems. The development of fair IR measures will provide tools to better inform policy makers about such issues and help practitioners to adhere to ethical, legal, and policy obligations established by governments.

2.1.3 Proposed research directions

To define what is fair is an ethical issue and a complex, context-dependent problem. As there is no one-fits-all definition, we aim to illustrate the different contexts in which being fair is crucial and their respective challenges. Fairness is domain dependent because what is fair in one domain may not be fair in another. For example, when presenting information about political parties during election time, we may consider it fair to provide an equal opportunity to each party to present their political agenda. However, when presenting scientific information, we may consider fair to provide always first the most updated version of some requested theory.

Fairness is time dependent because notions of ethics and fairness change as society changes (e.g., discrimination of minorities that in the past was not seen as a problem). Fair policies may embed a time variable in their formulation as they are required to change overtime in order to adjust for historical discrimination, e.g., obligatory quotas for historically discriminated groups.

Fairness is stakeholder dependent because what is fair for one stakeholder may not be fair for another. For example, in e-commerce services, customers would like to have recommended the best products with the lowest price, the e-commerce owners would like to sell the best products with the highest margins, while the producers would like to sell their product against the competition. Moreover, stakeholders may have various levels of abstractions, they can focus on an individual, group of people, or the society. Each of these levels may demand different definitions of fairness which may pose different trade-offs.

One of the main challenges in defining fair-IR measures is to include part, if not all, of these dimensions into their definitions. Other challenges include: the discovering and identification of biases and discrimination in IR systems; the definition of test collections based on identified and established real use-cases; the quantification of the impact of a fairness aware IR system; the identification of the risks and values of fairness, like the analysis of the trade-off between differential treatment versus disparate impact in various contexts such as hiring, lending, criminal justice system, etc.

2.1.4 Key challenges and obstacles

Fairness-aware measures are inherently related to the definition of fairness, which is acknowledged to be a multidisciplinary problem. Different parties will need to be involved in the process of identifying different definitions of fairness and making those definitions operational through computable measurements. We envisage collaborations between computer scientists, policymakers, social scientists, and lawyers, among others, to overcome this challenge.

Fairness can be seen as a desirable aspect of IR systems, which is complementary to other aspects typically considered in their evaluation, such as effectiveness or efficiency. How to combine fairness-aware measures with other evaluation dimensions is, however, still an under-explored problem.

As in other evaluation dimensions (e.g., there is no consensus on a single metric to evaluate effectiveness of an IR system), it is expected to see a variety of fairness measures. We will therefore need meta-evaluation frameworks to help researchers and practitioners in this area identifying which set of measures is the most appropriate to be used in a given scenario (e.g., optimizing fairness for a group versus optimizing fairness for each individual).

2.1.5 Impact

The development of tools, as measures and protocols of analysis, will assist researchers and practitioners in the detection and mitigation of biases and discrimination in IR systems. The detection and mitigation of biases in IR systems will positively impact the expected experience of under-represented or discriminated communities.

This research will inform policy makers to determine (1) policies and procedures to guarantee fairness of algorithms (including IR models) and (2) controls to support the implementation of these policies and procedures. It can also lead to improved decision making.

2.2 Build Fair-IR

In addition to determining which are the competing definitions and metrics for fairness in IR and selecting those that are most appropriate for a given scenario, building Fair-IR systems also requires an understanding of when to intervene, how to intervene, what information to provide to users, as well as of key trade-offs that need to be made.

2.2.1 Why is it important?

IR systems are the primary tools people use to access information. Hence, they could have a significant effect in changing, for instance, the way people think and the decisions they make; reshaping the society as a result. These aspects are also discussed in more detail in §2.1.2.

2.2.2 How is this relevant to IR?

Information retrieval, as a field, focuses on retrieving, ranking, understanding and presenting information to end users, with a potentially wide-raging impact. Therefore, as also argued in the prior subsection §2.1, fairness should be a primary focus when building IR systems.

2.2.3 Proposed research directions

When attempting to answer how one can build Fair-IR systems, we argue that there are a few key requirements and questions that have to be addressed, including: (1) The need for new, improved rankings that are based on fairer algorithms, which raises unique challenges that are specific to ranking, and requires understanding which protected attributes might be query dependent; (2) How to determine what metrics to optimise for, in order to achieve fairness (in terms of a variety of measurements like equal exposure or diversity) with respect to both the users for which the content is being surfaced and rendered, as well as the items or web pages that are being ranked? It is also important here to scrutinize if the metrics are and remain fair in expectation, given that, for instance, protected attributes may change over time. (3) Establishing when and where intervention should take place in an IR system pipeline (pre-processing versus in-processing versus post-processing), including how to account for complex IR systems that consist of multiple subsystems (fairness in ranking versus re-ranking), and how to account for often limited training data with respect to sensitive attributes. (4) The need for fairer user interfaces for IR systems that might need to balance between transparency versus fairness, an aspect that has been largely overlooked. (5) Understanding possible exploration versus exploitation trade-off, particularly in the case of personalization, including understanding how to minimize the reliance on "click" data that might be coming from unfair systems.

2.2.4 Key challenges and obstacles

Fairness in IR (and recommender systems) has received comparably less research attention than in classifier settings. In part, this appears to be due to the unique challenges it raises, on top of the challenges already present in more general machine learning fairness settings. For example, outputs are often a ranked list of limited size rather than independent binary actions, results are personalized which could touch upon demographics, the number of items to recommend is often very large, and the system affects multiple stakeholders (both consumers and producers). Likewise, there are also unique opportunities to show multiple items at a time, as well as the possibility of getting multiple opportunities to make recommendations over time. We believe understanding and exploring the unique properties of IR systems is crucial to making progress towards Fair IR. Development of fairness metrics must likewise take into account these unique facets of IR applications.

There are also numerous technical challenges in actually incorporating fairness in IR systems. One significant challenge is that these systems are typically trained on user feedback, which is often only observed over previous recommendations. Given the inherent sparsity of user feedback data over large item spaces, the designers of these systems should be mindful of their recall, as well as of possible explorations in order to be able to effectively evaluate and train future systems.

Second, there is the question of how to architect recommender and ranking systems to incorporate fairness. Traditionally, there has been a debate between using pre-processing

(modifying training data or representations), in-processing (model or loss function changes), or post-processing (changing results after model predictions). This is significantly complicated by the fact that production systems often are composed of numerous models and decision systems that together produce the end ranking. Where in this larger system to incorporate fairness mechanisms is an open question. Last, the system designer also controls the user interface by which users receive these recommendations, and should be mindful of how that interface can influence fairness.

There could also be limitations due to data availability. For example, we often do not directly observe the sensitive attributes, and this can be exacerbated by some of these attributes being contextual in nature. Knowing the sensitive attributes might be sometimes possible during training, but this is even more rare while serving users. In offline evaluation and testing on public datasets, we might also be limited by data collection biases (often from a previous IR system) and thus cannot properly evaluate how a system would perform if deployed in the future.

More philosophically, because there are often multiple stakeholders interacting with the IR system, user fairness and item fairness may be in tension. This creates a challenge of how to prioritize one stakeholder or another.

2.2.5 Impact

Publicly, a fair IR system enhances social fairness in an ethical viewpoint. First, by increasing the exposure of under-represented groups, it encourages opportunities of such groups. Second, it can also be useful for avoiding social problems, such as filter bubbles or the echo chambers (Resnick et al., 2013; Flaxman et al., 2016).

A fair IR system can also arbitrate between different stakeholders. For example, in job-searching scenario, job-applicants should be equally exposed in the search results returned to employers, irrelevant to their sensitive attributes.

2.3 Information Leakage Scenarios

Our discussion on information leakage scenarios focuses on the identification of, and protection against, content leakage, i.e., the leakage of sensitive or confidential information that is supplied by a content provider. It does not address, for example, how to secure the information about users that is captured by the system's log files through the course of using the system. Information leakage can occur both at the system level or through a user's actions as part of the broader system. Safeguarding against information leakage typically consists of:

(1) Identifying what needs to be protected, what is content and what are the levels of protection, including:
   (a) entire collections, entire documents or passages of text;
   (b) models or aspects thereof; and
   (c) known unknowns.
(2) Formalizing how to protect sensitive or confidential information, e.g., through
   (a) technology-assisted help for humans to identify sensitive content;
   (b) the implementation of confidentiality-aware end user search;
   (c) private information retrieval protocols;

(d) differential privacy for test collection release or multi-level security (such as role based access);
(e) cleaning up human leakage (such as finding all copies of leaked information); and
(f) mosaicing protection (protect against inference).

(3) Characterizing the limits of the protection, e.g.,
(a) designing an evaluation framework for assessing the level of protection, e.g. penetration testing, red team-blue team, etc.; and
(b) learning from other fields such as cybersecurity, game theory, and spam filtering.

(4) Learning from possible mistakes, e.g., to identify best practices.

2.3.1 Why is it important?

The IR community has made important progress in developing technologies that allow us to gain access to, aggregate, analyse and infer information and meaning about people, events, organizations, etc. However, the success of modern day search engines and IR technologies more broadly, is also limiting the amount of information that is made available to be searched. Search engines typically assume that all material that is available to them should be findable. However, an important fraction of the potentially storable words generated in the world in a day are generated in settings in which it simply is not practical to segregate appropriately findable content from sensitive content that requires protection. Examples include personal email, corporate intranets, communication between government entities, recorded teleconferences, and words spoken in the presence of a recording device (such as mobile phones or smart IoT devices).

As a result, document collections that contain potentially sensitive information are either not made available to the public at all or they require an expensive and time-consuming review to identify and protect all of the sensitive information before the collection can be considered for release. Developing IR technologies to identify and protect sensitive information, and methods for identifying and combating the leakage of such sensitive information, will enable a greater level of access to collections that have a risk of containing sensitive information. The IR community may ultimately approach a tipping point at which tasks that can be investigated without addressing concerns regarding sensitive information would become the minority. Indeed, some more recent IR tasks deal almost exclusively with sensitive content—for example, consider eDiscovery and technology-assisted sensitivity review (Grossman and Cormack, 2010; Oard et al., 2010; McDonald et al., 2018).

2.3.2 How is this relevant to IR?

There are at least four ways in which the information retrieval community can make essential contributions in this space. At the most basic level, identifying sensitive information can be cast as a classification problem; and text classification, in particular, has been of longstanding interest to information retrieval researchers. At a higher level, modern information retrieval systems are designed to optimize performance on some given task, and finding relevant content that is not sensitive among other sensitive content (and vice versa) can be seen as such a task.

At an even higher level, information retrieval addresses challenging tasks that exceed the capabilities of machines alone through the design of synergistic interactive "systems" that leverage both human and computing abilities. Automatically classifying and removing sensitive information from document collections such that the collections can be safely searched,

or responsively searching for disclosable content that is intermixed with sensitive content, are two examples of tasks that are within today's digital landscape and are becoming increasingly necessary.

At the highest level, information retrieval seeks to help people satisfy their information needs, and strong demand signals are already evident for some tasks, such as searching for evidence in lawsuits (i.e., "eDiscovery") and in governmental sensitivity review (e.g., to satisfy "freedom of information" obligations). Additional personal, corporate, and societal applications are likely to emerge over the coming years.

Sensitive information can exist within many content scopes that an IR system often need to process, e.g., individual passages of text, whole documents or entire collections. Moreover, sensitive information might not be explicit, it might be revealed by the ranking or classification models that comprise an IR system, or aspects of those models such as generated log data. There are, therefore, many new IR challenges that are to be addressed to enable the identification, and appropriate handling, of sensitive information for each of these content scopes.

2.3.3 Proposed research directions

We can identify four main lines of work that together would enable the creation of systems that can effectively handle sensitive information. The first is to determine what must be protected. This might be a document, a part of a document, or an inference that could be made given access to (parts of) several documents. Of these, the inference problem is the one that would be the greatest "stretch goal" for information retrieval research.

The second line of work would be on architectures for managing the search process. Protect-then-search and search-then-protect architectures are already in use for specialized tasks and such architectures can provide useful starting points, but there may also be advantages to jointly modeling relevance and sensitivity in some settings. It is worth noting that we must protect sensitive content not just from disclosure to the searcher, but also from disclosure to the search engine. This later task is the domain of the currently active research area known as "Private Information Retrieval."

The third line of work is to characterize the limits of the protection that can be afforded to sensitive content. Here there is a long heritage of evaluation-driven information retrieval research, but new evaluation measures, and perhaps also new approaches to evaluation, will be needed. For example, there may be requirements on the amount or nature of sensitive information for which disclosure could be risked, or quantification of how much information might be leaked in a worst case scenario.

Finally, when the information retrieval research meets practice, there will also be research needed on the design of policy frameworks that can help balance risks and benefits in ways that best leverage evolving technical capabilities, as well as on how best to identify and mitigate the consequences of unmodeled phenomena that could result in the unintended disclosure of sensitive content.

2.3.4 Key challenges and obstacles

Perhaps the most fundamental challenge will be the identification of sensitive and/or confidential content. Sensitivity is often not merely topical. Just as information retrieval research has evolved beyond a sole focus on topicality to encompass information characteristics such as freshness, veracity and intelligibility, sensitivity classification needs to be able to identify

characteristics of information beyond the well codified criteria of sensitive information categories used in fields such as medicine or financial data. Sensitivity classification needs to respect personal and contextual definitions of sensitivity that distinguish among users with different search goals, and the context in which the information was created. For example, information that is protected through freedom of information laws is often defined by broad-ranging categories with descriptions such as "information that would damage the international relations of a country," and making such a determination calls for understanding the source of the information, the context in which it was created or in which it is expected to be made available, and other nuances.

Adversarial behaviour, including collusion among multiple adversaries, will need to be addressed in some settings. For example, evaluation methods (such as algorithm deposit) that are suitable for use with truly sensitive content will need to provide assurances that facts about the collection can not be signalled through patterns in the evaluation results. As noted above, inference risks (which have been referred to as "information mosaicing" e.g., (Pozen, 2005)) will also demand consideration in some settings. Attention to the human element will be important as well. In particular, transparency and accountability will be key enablers for the adoption of new technical capabilities as they emerge. For example, in certain contexts, sensitivity classifiers need to align with legal and ethical frameworks to ensure that any actions that result from sensitivity classification predictions are lawful and fair.

2.3.5 Impact

The potential for impact extends well beyond the tasks such as privilege review in eDiscovery or sensitivity review for government documents that have been "early adopters" of this new technology. There are many user groups that could benefit from research on sensitive information leakage. For example, social scientists, historians and journalists are impacted by current access restrictions that are placed on document collections that have a sensitivity risk. Addressing the IR challenges associated with sensitive information could alleviate some of these restrictions and, in doing so, benefit society as a whole. Moreover, being able to provide a level of confidence that IR systems can effectively identify and handle sensitive information appropriately would positively impact individuals or organisations who would be adversely affected by the inadvertent release of their sensitive information.

In summary, by better protecting sensitive content we can, somewhat paradoxically, make more information available and searchable. The World Wide Web has shown us that if we make it possible for people to find the information they want, content providers will beat a path to our door. All that remains to do if we wish to unlock the potential of intermixed content (some of which is sensitive and some of which is valuable) or provide robust mechanisms for publicly releasing 'cleaned' versions of document collections that contain intermixed content at creation time, is to develop IR systems that can effectively identify and protect sensitive content while providing access to the content that is not sensitive; and convince potential content providers that we are able to do so.

2.4 Data Biases in Sensitive Information

Information retrieval research has also long dealt with various issues in the quality and availability of data; the complexity of these challenges only increase with the kinds of data

needed to study and audit IR systems with respect to social concerns, including fairness, accountability, and safety. One reason is the increased sensitivity of some of this data. Studying fairness, for example, often requires sensitive information such as membership in protected classes (gender, race, religion, etc.) in order to identify discrimination against particular groups.

Key challenges and questions include: (1) How do we educate student researchers, computational modelers, public data users, and other relevant stakeholders about data issues? (2) How do we evaluate the sensitive data itself? (3) How do we encourage and support research on such data? (4) How do we safeguard, release, and use sensitive data? (5) How do we encourage social science grounding of data? (6) How do we use diverse forms of data effectively in study design?

2.4.1 Why is it important?

Biases in sensitive data can affect different demographics and stakeholders differently. If sensitive data is collected or available at higher rates for certain groups of users (e.g., medical data about adult cancer patients), that may put them at greater risk in case of leakage or misplacement. Conversely, due to privacy concerns sensitive data might be collected at lower rates for other groups of users (e.g., medical data about children with cancer), which may result in sub-optimal decision making in some application domains. Such biases are also critical in the context of the debate about audits and fair algorithms that could require in certain instances access to such sensitive information in order to assess if, for instance, a system offers the same quality of service for users with different characteristics. Data and systems can have a circular feedback loop, where bias in the data results in biased system performance, which induces further biases in the data collected from its users. Understanding biases in the context of IR data in general, and sensitive information in particular, is thus critical.

2.4.2 How is this relevant to IR?

People are applying fairness and bias approaches to IR problems now, or using IR approaches in ways that impact the world in more far reaching ways than have been previously considered. As emphasized throughout this section, IR technologies are being used far beyond their original scope and we have a responsibility to ensure that these applications are not exacerbating existing social problems or creating new ones as a result of their data, particularly unexamined biases that this data may contain.

Acquiring data is a crucial part to evaluating IR systems: if evaluation data has biases, the evaluation will be flawed—data driven analysis is king! Sensitive scenarios in which we use past user behavior and query logs may surface this data in ways that can be problematic (leakage), and how we leak may reflect data biases or have more harmful consequences. All this adds additional dimensions to the issues the IR community already wrestles with, especially issues that are fraught ethically and legally. Given IR's focus on evaluation, we would benefit from a more nuanced understanding of the many ways bias and artificiality[1] can enter in our work.

[1]We use the term *artificiality* to distinguish between data reflecting naturally occurring phenomena and data shaped, manipulated or artificially produced by humans or systems with a given goal.

2.4.3 Proposed research directions

There are several key research directions that we believe the community should consider. One direction regards the evaluation of sensitive data, whether that data is found, synthetic, or a mix of both. Specifically, for synthetic and found data collections, and the "semi-synthetic" middle spectrum, how do we evaluate these collections and assess their biases and the subsequent impact on our work? This applies both to research work — running studies to understand system behavior — and to the development and deployment of the IR systems themselves.

Another question is to understand protection schemes for sensitive data: beyond best practices for effective custodial care and handling, how do we encourage people to work on these important problems and datasets, and yet effectively understand and appreciate the risks? How can the community nurture and support this work rather than just identifying and penalizing errors?

We also need a deeper and more nuanced understanding, with taxonomies and theories, of the many diverse ways bias or artificiality can enter in our study or system designs (Olteanu et al., 2019a; Baeza-Yates, 2018).

The field also needs research on how to proactively address potential harms that we may not yet know how to handle or even identify. Research on fairness, accountability, and transparency is building an ever-expanding set of tools for measuring and correcting biases that we know to look for, but techniques for preempting potential current and future biases that may not yet be on the community's radar are not nearly as well developed or understood. Thus, current efforts that ground the discovery and exploration of biases in prior literature in social sciences (including psychology, linguistics, sociology, and economics), are critical to identifying new biases and mechanisms to quantify and track them, and should be expanded.

Finally, we also believe educational efforts are critical here, specifically how to effectively educate both the community and the public about these issues. One direction could be efforts similar with the translation or application tutorials at FAT* that aim to explain issues around data biases in the context of sensitive information, and that expose participants to relevant case studies. In any case, we need training to teach IR researchers best practices for using such data in their research; practitioner's methods for measuring and mitigating the effects of data biases on the systems they build; and to inform the public in understanding more accurately the ways in which IR systems, their underlying data, and society interact.

2.4.4 Key challenges and obstacles

There are quite a few challenges that this agenda needs to overcome. These include:

(1) issues with *accessing data* (that might be sensitive, confidential, about protected groups, and proprietary) in order to support research;
(2) providing adequate *education*, such as for students (especially machine learning) without experience in collecting and analyzing sensitive data;
(3) *data issues* (real, synthetic, and everything in-between) in study design have important effects on research and practical impact on IR systems, yet are messy, complicated, and often unappreciated by researchers, reviewers, and funders; and
(4) *misrepresentation* in data collections likely has a disproportionate impact on groups who tend to be already underrepresented and marginalized, being reflective of existing societal exclusions.

2.4.5 Impact

We believe more research in this area will provide the basis for new and improved guidelines for the next generation of researchers in their increased work with various sorts of sensitive data.

Users of IR systems, and society in general, will benefit from better information science and the ability to build better systems that are less influenced by the biases that creep from society through their training and evaluation data. We will be able to more accurately hold the developers and operators of IR tools accountable for the effects of their system because of more accurate data and methods for assessing those impacts, while implementing appropriate safeguards for handling sensitive data.

Finally, a more nuanced and accurate understanding of the validity and bias issues in the selection and use of data, particularly sensitive data, will improve the quality (and hopefully quantity) of research in this space.

2.5 Build Transparent-IR

When attempting to build transparent-IR, the key questions are: What is the purpose of transparency? Can we fundamentally redefine methods to make them transparent without sacrificing efficiency and violating user privacy? How can we quantitatively evaluate transparency both intrinsically and extrinsically?

2.5.1 Why is it important?

Currently, there is huge distrust of automated decision making systems in society. Transparency is a way to repair the growing distrust, and can be achieved in different ways, including through providing explanations as to why a system gave a certain recommendation, or by making explicit to the end-users what type of data the system is collecting and/or using for these recommendations. With effective transparency, systems will become auditable and users will be able to find errors, as well as learn how to use these systems better in order to obtain the content they seek.

There are also growing legal structures (e.g., General Data Protection Regulation in Europe, known as GDPR) whose impact on automated systems needs to be researched.

2.5.2 How is this relevant to IR?

IR systems are globally used tools accessed by the majority of the world's population. Their impact is only growing further with the increasing proliferation of systems that rank objects relative to some form of user need (e.g. recommender systems).

The field of information retrieval has decided what documents users see for over 20 years. This gives a huge test case of the impact of algorithmic decision making on society and how transparency may help. Did IR foster fake news?

2.5.3 Proposed research directions

We need to determine the purpose of transparency: for instance, in some application areas, the purpose is to be actionable, to allow users to change the outcome; in others, it is knowing

why, which might affect the system down the line. For users, the purpose of transparency may vary based on context or domain, or we may even require personalised transparency.

We believe there is a need to create models of transparency and develop methods to present them to the users. The models of transparency may vary depending on the purpose of transparency, and their presentation method is dependent on the modality of the system (e.g., text-based versus audio-only search systems).

Both intrinsic and extrinsic evaluation methodologies need to be established in order to test the effectiveness of systems that support transparency. Quantifiable metrics will need to be defined in order to determine the utility of the methods and also facilitate reproducibility.

2.5.4 Key challenges and obstacles

Requiring genuine transparency may limit the array of methods that can or cannot be used (e.g., neural approaches to learning). Other downsides of transparency might include trade-offs, lower accuracy, slower algorithms, and, consequently, lower trust in the system or decreased overall user satisfaction.

Businesses do not always have the incentives to be fully transparent about the intricacies of their systems, as that might negatively impact their business. In such a case, the challenge is to identify ways in which various models of transparency can be beneficial to their business.

Quantifying transparency is also a major challenge, and evaluating transparency and its impact will likely need to be task and domain specific.

Finally, privacy might be an obstacle, as there might a tension between ensuring privacy and providing explanations. Some explanations may reveal too much personal information and conflict, for instance, with the right to be forgotten (Rosen, 2011; Veale et al., 2018a,b).

2.5.5 Impact

The extent of IR and recommender systems means that allowing transparency will empower both users and content providers to understand how the systems that they use operate. It would not be an overstatement to say that this will and has already affected the world. Organizations that operate such systems will be trusted more, as these organizations will become accountable.

The realization of such work can only occur through the collaboration of computer science and other research fields. The act of collaborating will itself have an impact on the individual research fields, transforming them into social-technical research domains.

2.6 Transparency Dimensions

Obtaining transparency in information retrieval research and applications could benefit from building a stronger theoretical and methodological foundation. Several transparency dimensions could be explored and some of the core challenges in achieving this goal are summarised below.

2.6.1 Why is it important?

Transparency is important because people should be informed about the mechanisms underlying the automated consequential decision-making that impacts their everyday lives. Furthermore, transparency is necessary in order to address other ethical considerations, such as

fairness and accountability. Research on transparency in IR has a wide range of motivations and applications; hence the need for academic consensus on the dimensions of transparency.

2.6.2 How is this relevant to IR?

As highlighted earlier, information retrieval is about connecting people to information. This should include information about the decision-making process of information retrieval systems. Additionally, the insights provided by increased transparency of our IR systems could help to further improve them.

2.6.3 Proposed research directions

An important challenge is to strengthen the theoretical and methodological foundation of research on transparency in IR. In order to build a theoretical foundation, the dimensions of transparency need to be explored. For this, a key question is how to formalize definitions, components and criteria for transparency. Contributions to the methodological foundation should focus on methods and metrics for evaluating transparency. We are not necessarily advocating for a universal approach – given the variety of use cases we recognize this might not be feasible. Nonetheless, it would be useful to have a more concrete set of approaches that can be applied in various settings.

In addition to these main challenges, we need to bear in mind the intention is to provide humans with insights. User experience should be taken into account to assure transparency in IR is useful and relevant. This includes both the user experience of end users and developers, and entails assessing usefulness and relevance for multiple applications of transparency (e.g., promoting user trust, discovering bias, generating explanations, improving systems).

Open research questions include:

- How can we formalize the definitions, components and criteria for transparency?
- What methods and metrics can be developed for evaluating transparency?
- How can we design user experiments to most effectively evaluate transparency?
- How can you promote transparency in all steps of the development and deployment pipeline (e.g., intentions of data collection, defining the constraints of problem)?
- How can you use agency to drive transparency, and transparency to drive agency?
- How can you give agency to both the users and creators of deployed systems in order to drive transparency?
- How can you use transparency in order to provide agency to users of systems so they can make informed decisions about how they interact with these systems?
- What are the most effective algorithms for generating various types of explanations?
- How does the way you present an explanation affect the user experience of transparency?
- What kind of explanations best promote user trust in a system?
- How can transparency be used to ensure ethical considerations of a system are adequate (e.g., discovering unfairness or bias, understand which inputs led to an observed output)?

2.6.4 Key challenges and obstacles

There are two main challenges. The first is the formalization of the topic of research; this includes consensus on the definitions, components and criteria for transparency in information

retrieval systems. The second is in evaluation, where we encourage contributions on methods and metrics for evaluating transparency for various users and applications.

2.6.5 Impact

The overall field of research on transparency in IR will mature from contributions to the theoretical and methodological foundations. Transparency of information retrieval systems will impact anyone who builds, deploys or uses IR systems. It can provide insight in algorithmic decision-making, increase user trust, and give control to the users that are affected by the system. It can also help developers query for weak spots, avoid potential harm due to malfunction, manipulation or bias, and comply to regulations. Lastly, it can bolster research on information retrieval by increasing the understanding of our systems.

2.7 Harms in IR Systems

We believe that deliberations about harms in IR systems should delve into several key issues, including: (1) what is harm and what types of harm exist in IR systems? (2) what is the relation between harmfulness and other topics in the FACTS space? (3) how can we measure harm in a generalizable way? (4) what is ethical and legal in terms of interventions in this problem space? (5) when trying to prevent harms, how much operational freedom should the system have and how much agency should it leave to its users?

2.7.1 Why is it important?

Many users interact with information systems daily. While the benefits of broad access to knowledge and information are undeniable, we have been observing evidence of harmful offline impact that online systems have on societies as well as individuals and their mental and physical well-being. By now, research has uncovered a number of problems that fall under this umbrella. Those problems include but are not limited to misinformation, disinformation, public opinion manipulation, technology addiction, or exposure to sensitive or objectionable content such as violence, aggression or pornography.

2.7.2 How is this relevant to IR?

IR systems mediate access to information and thus have the power to shape people's knowledge and beliefs. This knowledge and beliefs in turn indirectly impact people's behaviors and actions. Such influence over their users makes IR systems a potential source of safety and harm concerns discussed in the previous section. Exposing users to harmful content could diminish their trust and as a result inhibit future acceptance and deployment of IR technology.

2.7.3 Proposed research directions

There are two important and relevant research themes. The first theme concerns the understanding and measurement of harm. We need to understand what types of harms exist in IR systems, what constitutes harmful content, and how it may influence users in the real world. Once harms are understood, we moreover need to design reliable and generalizable measurement techniques allowing us to trace the effects of harmful content outside of IR

systems in various contexts and domains. Lastly, we think it is crucial to better understand the interplay between harmfulness and other problems in the FACTS-IR space. For instance, can search results be ever harmful but fair? Reversely, can they be harmless but unfair?

The second research theme concerns the ethics and legality of interventions in this problem space. To what extent can and should an IR system results be modified to prevent harms? Who should decide this and how much agency should users have? How do we make sure the interventions do not cause new and unpredicted harms?

2.7.4 Key challenges and obstacles

We expect the key challenges in this research area to closely follow the most exciting research directions and revolve around the difficulties of pinpointing and measuring harms, as well as the ethical and legal constraints. We believe harms are highly contextual and dependent on the specifics of a given domain and life situations of the individuals involved. This contextual and individual complexity will not only directly contribute to the challenge of understanding harms, but also to the potential difficulty of gaining wide public acceptance of the proposed solutions. Accountability and data tracing in the context of harm analysis might lead to potential tensions with regulations such as GDPR, which requires explicit user consent for new uses of data. Furthermore, mitigating harms in high-stakes domains, such as politics, will expose us to hard normative choices which should not be made by technologists themselves but in collaboration with domain experts and ethicists.

Last but not least, measurement of real-world harms will face the obstacle of a proper experimental design. Monitoring harms might require conducting longitudinal studies with all their associated challenges, such as finding adequate participants and controlling for plethora of confounding factors. The latter obstacle will also play a role in studies that try to correlate offline harms with IR system interactions, given that negative effects might manifest themselves long after a user is exposed to an IR system.

2.7.5 Impact

We predict that work in this space will have an impact on multiple stakeholders. First, those consuming the results produced by IR systems are influenced by the information they are exposed to. For instance, search results in the health domain may cause harm when a user searching for symptoms and conditions is not exposed to all the relevant information and as a result decides not to seek a necessary treatment. Reversely, exposure to too much information in the health domain may trigger unnecessary anxiety. In the political space, a skewed information consumption makes it easy for adversarial actors to manipulate their audiences. While these effects may occur a long time after a user interacts with a system, we believe that IR systems may also cause harms in more immediate ways. For instance, a badly designed interface with too bright, flashy content may trigger seizures in people with epilepsy, while exposure to certain individually traumatic content may negatively influence people with a post-traumatic stress disorder (imagine a person who has just lost their newborn child seeing a diaper ad.)

The second group we envision being impacted by this research are people presented in the results of IR systems. For instance, certain disadvantaged or minority groups, as well as individuals, might be misrepresented or depicted in a biased way in documents or images surfaced by search engines. The consequences of such misrepresentation might extend to the offline world, leading to encouragement and reinforcement of existing prejudices, or even to

violent physical targeting of people in the most extreme cases. Understanding and detection of such misrepresentation might help prevent these offline harms.

Last but not least, research in this space will impact those who contribute to the development of IR systems. Examples include content moderators or data annotators who have to manually sort through harmful content (Karunakaran and Ramakrishan, 2019; Dang et al., 2018). Understanding of what constitutes harmful content is the first necessary step to develop accurate automated methods to detect it. Furthermore, in addition to considering the emotional impacts of moderation work, we must treat demographics with particular care due to potential impacts findings could have on employment. If certain demographic groups are more impacted by moderation work, sharing that could lead to hiring discrimination against those groups – thus, attempting to reduce one form of harm could lead to another. Understanding who are the people least harmed and how to best portion and divide the potentially sensitive content among the system contributors could help limit the negative effects.

3 Parting Thoughts

The information retrieval community has the responsibility to care about the broader impact and implications of the systems that we research and the systems that we build in academia and industry. This responsibility is articulated, among other places, in the ACM Code of Ethics, which includes a responsibility to be proactive about identifying and preventing potential harms that may arise from our work, even when the intentions of that work are beneficial. Similar responsibility issues are also being addressed in related fields, with, for instance, the emergence of the community around the ACM Conference on Fairness, Accountability, and Transparency (see `https://fatconference.org/`), a venue with a cross-disciplinary focus that brings together a diversity of researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

However, there are specific issues in IR stemming from the characteristics of and the reliance on document collections, and the often imprecise nature of search and recommendation tasks. IR has a strong history of using test collections during evaluation, but the biases built into these collections and their surrounding evaluation protocols are not fully understood, particularly biases related to historical and ongoing societal inequities. For example, the people who construct topics and make relevance assessments arguably are not necessarily representative of the larger population. In some cases, they have not been representative of the type of users who are being modeled (e.g., having people who do not read blogs evaluate blogs). Evaluation measures are also designed to optimize certain performance criteria and not others, and either implicitly or explicitly have built-in user models. Systems are then tested and tuned within this evaluation framework, further reinforcing and reifying any existing biases (Allan et al., 2018). Safety and privacy issues are also prevalent within most IR applications, as they tend to record vast information about their users and are sometimes prone to manipulation for business or political purposes.

Given the central role that IR technology plays in today's society, it is critical to continue to build a community of researchers and practitioners to characterize and address FACTS-related issues. The agenda setting activities of this workshop were meant to do just that.

Author details and funding:
- Adam Roegiest, Kira Systems, adam.roegiest@kirasystems.com.

- Aldo Lipani, University College London, aldo.lipani@ucl.ac.uk.
- Alex Beutel, Google, alexbeutel@google.com.
- Alexandra Olteanu, Microsoft Research, alexandra.olteanu@microsoft.com.
- Ana Lucic, University of Amsterdam, a.lucic@uva.nl, was supported by the Netherlands Organisation for Scientific Research (NWO).
- Ana-Andreea Stoica, Columbia University, astoica@cs.columbia.edu.
- Anubrata Das, University of Texas at Austin, anubrata@utexas.edu, was supported by the Knight Foundation, the Micron Foundation, and Wipro.
- Asia J. Biega, Microsoft Research Montréal, asia.biega@microsoft.com.
- Bart Voorn, Ahold Delhaize, bart.voorn@aholddelhaize.com.
- Claudia Hauff, Delft University of Technology, c.hauff@tudelft.nl, was supported by the Netherlands Organisation for Scientific Research (NWO).
- Damiano Spina, RMIT University, damiano.spina@rmit.edu.au.
- David Lewis, Brainspace Corporation, davelewis@daviddlewis.com.
- Douglas W. Oard, University of Maryland, oard@umd.edu.
- Emine Yilmaz, University College London, emine.yilmaz@ucl.ac.uk.
- Faegheh Hasibi, Radboud University, f.hasibi@cs.ru.nl.
- Gabriella Kazai, Microsoft, gkazai@microsoft.com.
- Graham McDonald, University of Glasgow, graham.mcdonald@glasgow.ac.uk, was supported by the EPSRC IAA programme scheme.
- Hinda Haned, Ahold Delhaize and University of Amsterdam, hinda.haned@aholddelhaize.com.
- Iadh Ounis, University of Glasgow, iadh.ounis@glasgow.ac.uk, was supported by the EPSRC IAA programme scheme.
- Ilse van der Linden, University of Amsterdam, iwc.vanderlinden@gmail.com.
- Jean Garcia-Gathright, Spotify, jean@spotify.com.
- Joris Baan, University of Amsterdam and DPG Media, jsbaan@gmail.com.
- Kamuela N. Lau, Georgia Institute of Technology, klau43@gatech.edu.
- Krisztian Balog, University of Stavanger and Google, krisztian.balog@uis.no.
- Maarten de Rijke, University of Amsterdam, derijke@uva.nl, was supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), and the Innovation Center for Artificial Intelligence (ICAI).
- Mahmoud F. Sayed, University of Maryland, mfayoub@cs.umd.edu, was supported by NSF award 1618695.
- Maria Panteli, BBC, m.x.panteli@gmail.com.
- Mark Sanderson, RMIT University, mark.sanderson@rmit.edu.au.
- Matthew Lease, University of Texas at Austin, ml@utexas.edu, was supported by the Knight Foundation, the Micron Foundation, and Wipro.
- Michael D. Ekstrand, Boise State University, michaelekstrand@boisestate.edu, was supported by NSF award 1751278.
- Preethi Lahoti, Max Planck Institute for Informatics, plahoti@mpi-inf.mpg.de.
- Toshihiro Kamishima, National Institute of Advanced Industrial Science and Technology (AIST), mail@kamishima.net.

References

James Allan, Jaime Arguello, Leif Azzopardi, Peter Bailey, Tim Baldwin, Krisztian Balog, Hannah Bast, Nick Belkin, Klaus Berberich, Bodo von Billerbeck, Jamie Callan, Rob Capra, Mark Carman, Ben Carterette, Charles L. A. Clarke, Kevyn Collins-Thompson, Nick Craswell, W. Bruce Croft, J. Shane Culpepper, Jeff Dalton, Gianluca Demartini, Fernado Diaz, Laura Dietz, Susan Dumais, Carsten Eickhoff, Nicola Ferro, Norbert Fuhr, Shlomo Geva, Claudia Hauff, David Hawking, Hideo Joho, Gareth Jones, Jaap Kamps,

Noriko Kando, Diane Kelly, Jaewon Kim, Julia Kiseleva, Yiqun Liu, Xiaolu Lu, Stefano Mizzaro, Alistair Moffat, Jian-Yun Nie, Alexandra Olteanu, Iadh Ounis, Filip Radlinski, Maarten de Rijke, Mark Sanderson, Falk Scholer, Laurianne Sitbon, Mark Smucker, Ian Soboroff, Damiano Spina, Torsten Suel, James Thom, Paul Thomas, Andrew Trotman, Ellen Voorhees, Arjen P. de Vries, Emine Yilmaz, and Guido Zuccon. Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52:34–90, June 2018.

Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6), 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414. ACM, 2018.

Brandon Dang, Martin J. Riedl, and Matthew Lease. But who protects the moderators? the case of crowdsourced image moderation. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*, 2018.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128. ACM, 2019.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, 144:114–126, 2015.

EU. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, L119:1–88, 2016.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

Eric Goldman. Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology*, 8:188, 2005.

Michael Golebiewski and danah boyd. Data voids: Where missing data can easily be exploited. *Data & Society Research Institute*, 2018.

Maura R Grossman and Gordon V Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17:1, 2010.

Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1914–1933. ACM, 2017.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*2019)*, pages 49–58, 2019.

Lucas D Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *The Information Society*, 16(3):169–185, 2000.

Sowmya Karunakaran and Rashmi Ramakrishan. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 50–58, 2019.

Farhad Manjoo. Right to be forgotten online could spread. *New York Times*, 2015. URL `\url{https://www.nytimes.com/2015/08/06/technology/personaltech/right-to-be-forgotten-online-is-poised-to-spread.html}`.

Graham McDonald, Craig Macdonald, and Iadh Ounis. Active learning strategies for technology assisted sensitivity review. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 439–453. Springer, 2018.

Rishabh Mehrotra, Amit Sharma, Ashton Anderson, Fernando Diaz, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th International Conference on World Wide Web*, pages 626–633, 2017.

Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. Twitter's glass ceiling: The effect of perceived gender on online visibility. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law*, 18(4): 347–386, 2010.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019a.

Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, and Michael D. Ekstrand, editors. *Proceedings of FACTS-IR 2019*, 2019b. CoRR abs/1907.05755. URL `http://arxiv.org/abs/1907.05755`.

Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, and Michael D. Ekstrand. Workshop on fairness, accountability, confidentiality, transparency, and safety in information retrieval (FACTS-IR). In *SIGIR 2019: 42nd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1423–1425. ACM, July 2019c.

David E. Pozen. The mosaic theory, national security, and the freedom of information act. *The Yale Law Journal*, 115:628–679, 2005. Available at: `https://scholarship.law.columbia.edu/faculty_scholarship/573`.

Paul Resnick, R Kelly Garrett, Travis Kriplean, Sean A Munson, and Natalie Jomini Stroud. Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pages 95–100. ACM, 2013.

Jeffrey Rosen. The right to be forgotten. *Stanford Law Review*, 64:88, 2011.

Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM, 2011.

Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and Maarten de Rijke. Finding influential training samples for gradient boosted decision trees. In *Proceedings of the International Conference on Machine Learning*, page 4584–4592, 2018.

Amanda Spink and Michael Zimmer. *Web Search: Multidisciplinary Perspectives*, volume 14. Springer Science & Business Media, 2008.

Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pages 923–932. International World Wide Web Conferences Steering Committee, 2018.

Michael Veale, Reuben Binns, and Jef Ausloos. When data protection by design and data subject rights clash. *International Data Privacy Law*, 8(2):105–123, 2018a.

Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180083, 2018b.

Peng Wang, Xianghang Mi, Xiaojing Liao, XiaoFeng Wang, Kan Yuan, Feng Qian, and Raheem Beyah. Game of missuggestions: Semantic analysis of search-autocomplete manipulations. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2018.

Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. ACM, 2013.

Hui Yang, Ian Soboroff, Li Xiong, Charles L.A. Clarke, and Simson L. Garfinkel. Privacy-preserving IR 2016: Differential privacy, search, and social media. In *SIGIR*, pages 1247–1248. ACM, 2016.