A Systematic Review of Fairness, Accountability, Transparency and Ethics in Information Retrieval

NOLWENN BERNARD, University of Stavanger, Norway
KRISZTIAN BALOG, University of Stavanger, Norway

We live in an information society that strongly relies on information retrieval systems, such as search engines and conversational assistants. Consequently, the trustworthiness of these systems is of critical importance, and has attracted a significant research attention in recent years. In this work, we perform a systematic literature review of the field of fairness, accountability, transparency, and ethics in information retrieval. In particular, we investigate the definitions, approaches, and evaluation methodologies proposed to build trustworthy information retrieval systems. This review reveals the lack of standard definitions, arguably due to the multi-dimensional nature of the different notions. In terms of approaches, most of the work focuses on building either a fair or a transparent information retrieval system. As for evaluation, fairness is often assessed by means of automatic evaluation, while accountability and transparency are most commonly evaluated using audits and user studies. Based on the surveyed literature, we develop taxonomies of requirements for the different notions, and further use these taxonomies to propose practical definitions to quantify the degree to which an information retrieval system satisfies a given notion. Finally, we discuss challenges that have yet to be solved for information retrieval systems to be trustworthy.

CCS Concepts: • **Information systems** → **Information retrieval**; • **Social and professional topics**;

Additional Key Words and Phrases: Information Retrieval; Ethics; Fairness; Accountability; Transparency

1 INTRODUCTION

We live in an information society where we have grown to crucially depend on automated tools, such as search engines and conversational agents, that facilitate access to information. The research and development of these tools is the subject of the field of information retrieval (IR), which is defined as being concerned with "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [65]. IR systems are one of the most advanced and most widespread form of artificial intelligence (AI)—they aim to understand the *meaning* behind the user's query and respond appropriately. Specifically, in this work, we keep a narrow focus, where an IR system is defined to be one that receives a textual query expressing an information need and returns a ranked list of relevant items from a collection. At their core, IR systems boil down to the problem of *ranking* items based on their estimated relevance to the query. In the basic IR setting, it is further assumed that the system has no background information or historical behavior data about its user, i.e., this ranking is non-personalized. Early IR systems relied on ranking functions that capture the goodness of a match between a query and a document using various heuristics (e.g., TF-IDF weighting) [31]. The desire to combine multiple signals in the ranking function in a non-heuristic fashion has led to the development of *learning-to-rank* approaches in the 2000s [61]. There, various

Authors' addresses: Nolwenn Bernard, University of Stavanger, Stavanger, Norway, nolwenn.m.bernard@uis.no; Krisztian Balog, University of Stavanger, Stavanger, Norway.

ACM 0360-0300/2023/12-ART
https://doi.org/10.1145/3637211

intuitions of what makes a good match are captured in hand-crafted features and machine learning is employed to learn the optimal combination of these features based on training examples. Most recently, deep learning has transformed the field of IR as well, and an array of *neural IR* approaches have emerged that eliminate the need for manual feature design [71]. Looking back at several decades of progress, we can observe how the field has moved to more and more advanced forms of AI, which has consequently led to more and more effective systems. At the same time, these systems are becoming less and less transparent and increasingly more 'black box,' where even system designers may not fully understand how certain results are obtained. Of specific concern recently is the reliance on ever larger neural language models, which can produce output that is fluent and coherent in its own right, yet inaccurate [9]. Given the widespread use of information access systems, and our reliance on them as a society, their trustworthiness is of fundamental importance. It is essential to recognize that regardless of the continuous advancements in user interfaces and functionality, modern information access systems still address an IR ranking problem at their core. These systems employ multi-step processing pipelines, starting with initial stage retrieval, and encompass various applications such as search engines [104], recommender systems [2], and conversational assistants [60]. Thus, IR systems can impact users individually as well as society society at large. For example, Kay et al. [52] shows that users' perception of gender proportion in occupations can evolve after being exposed to manipulated search results. Thus, balancing gender proportion in search results might tackle stereotype exaggeration. In this work, we associate the notions of fairness, accountability, transparency, and ethics as requirements for an IR system to be trustworthy.

The challenges of fairness, accountability, transparency, and ethics (referred to as FATE henceforth) are not specific to information retrieval—they are shared by several sub-fields of artificial intelligence. These topics have been receiving a rapidly growing attention from the research community in the last years, as illustrated by the increasing number of publications. For example, the number of submissions to the ACM Conference on Fairness, Accountability, and Transparency (FAccT)[1] increased by around 350% between 2018 and 2021. The need for safe and fair use of AI has also been recognized in regulatory attempts such as the General Data Protection Regulation (GDPR)[2] and the Artificial Intelligence Act [78] in the European Union, the California Consumer Privacy Act (CCPA) in the United States,[3] or the Artificial Intelligence and Data Act Bill in Canada.[4] Concurrently, it is important to notice that FATE notions through the lens of IR are different from the FATE notions used to study machine learning (ML) in general. When people use information retrieval systems, they become the main driver of a human-machine interaction that happens in real time. There is an unusually large freedom to ask the system about virtually anything and get an answer. This is unlike many other ML applications (e.g., image classification or text clustering) where people are less directly affected by the system's predictions and where it is easier to introduce additional safeguards against mistakes and errors.

This work presents a *systematic literature review* [53] on the notions of fairness, accountability, transparency, and ethics in IR systems. While ours is not the first attempt to survey this field, there are two essential characteristics that differentiate it from existing overviews [7, 30]. First, our survey follows a systematic review protocol to provide a comprehensive synthesis of the available publications on FATE focusing on the core problem of ranking in IR from 1980 to the present day, as opposed to centering around the broader topic of information access [30]. By following a systematic approach, we also avoid a potential researcher bias. We initially retrieved 2,049 papers following a structured search process and selected 75 of them for inclusion in our review based on well-defined criteria. The complete list of papers considered and their annotations are made publicly available to support reproducibility. Second, our scope is not limited to a single FATE notion but considers all of them in depth. We also look at the interplay and possible tensions between the different notions. The objective of this survey is to

[1]https://facctconference.org/
[2]https://gdpr-info.eu
[3]https://www.oag.ca.gov/privacy/ccpa
[4]https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading

provide a general overview of the development regarding FATE in IR over the last decades, from their definitions and implementation to their evaluation. As part of this, we aim to bring clarity to the different dimensions and aspects related to the FATE notions in IR, synthesize key research results, and identify open challenges. Our main findings include the following:

- The different FATE notions do not have clear definitions. Therefore, we identify dimensions from the literature that contribute to the development of taxonomies of requirements for each notion. Having precise taxonomies aids the standardization and operationalization of these notions.
- Specifically, for fairness, transparency, and accountability, we identify three main dimensions and some related sub-dimensions. However, we can hardly single out dimensions for ethics. Indeed, ethics is concerned with other considerations such as privacy and safety which are distinct.
- Most of the work in the field has focused on either fairness or transparency. A significant number of studies on fairness or transparency tend be model agnostic and applied on the output of the retrieval algorithm. Although fairness can be considered at the individual and group levels, we notice that individual fairness remains largely unexplored. Furthermore, we observe that accountability is mostly mentioned in regulations rather than in the descriptions of the proposed algorithms or systems.
- Some tension and interplay can happen between different notions as well as within a given notion's dimensions. For example, a tension exists between individual and group fairness, i.e., optimizing one does not always imply the optimization of the other. Moreover, transparency can reduce the interpretability of the system if it is not adapted to the cognitive load of the users.

Based on the results of the literature review, we also identify remaining challenges, thereby providing directions for future research. The responsibility of different actors, development of regulations, and new evaluation methods are few examples.

In summary, the main contributions of this work are threefold. First, we perform a systematic literature review of FATE notions in IR systems from 1980 until now, and provide a broad and reasoned overview of the state of the art. Second, we take a step towards a standard formalization of the notions of fairness, accountability, transparency, and ethics based on the outcome of the review. Finally, we discuss open challenges related to the different notions of FATE in addition to raising the question of how to build IR systems that combines all of them.

2 SYSTEMATIC REVIEW PLANNING

A systematic literature review (also referred as systematic review) is a type of study that follows a well-defined methodology to curate, analyze, and synthesize relevant literature on a specific research question. Describing the methodology used makes the process transparent and supports reproducibility. In this work, we follow the guidelines presented by Kitchenham and Charters [53], and divide the systematic literature review in three main phases: planning, conducting, and reporting the review. In this section, we present the goals and the research questions that we seek to answer. Moreover, the search strategy is specified including the source databases and explicit selection criteria for inclusion and exclusion.

As mentioned above, the field of FATE in IR is attracting attention, however existing survey papers tend to either focus on a specific aspect of FATE or study a limited selection of papers. This systematic literature review aims to provide a comprehensive overview of the existing work in the field as well as its evolution over the last 40 years.

2.1 Research Questions

This work ought to help researchers in the field of IR to have a general idea of what has already been investigated over the last decades with regards to FATE. To get started, we need to understand how each FATE notion is defined in the context of IR and what are its associated characteristics. Once the definitions are established, we look at

the operationalization of these notions to build a trustworthy (i.e., fair, accountable, transparent, and ethical) IR system. Next, we move on to the evaluation of FATE notions that help to compare the different approaches proposed to create a trustworthy system. Finally, the conclusions reached in different studies and workshops to gain a better understanding of the field from both a technical and a societal point of view. Consequently, the following research questions are identified for the systematic review:

RQ1: What are the definitions and characteristics of fairness, accountability, transparency, and ethics in IR?
RQ2: How to build a fair, accountable, transparent, and ethical IR system?
RQ3: How are FATE notions evaluated?
RQ4: What conclusions emerge from foundational and empirical studies as well as from discussions at workshops?

Grounded on the findings of the review, we develop taxonomies based on the different dimensions (and sub-dimensions) identified for each FATE notion, along with remaining open challenges in the field.

2.2 Databases

Gusenbauer and Haddaway [43] provide a detailed analysis of 28 academic search systems to help researchers to select the most adequate scholarly databases when performing systematic reviews. Based on their analysis, the topic of this review, and additional considerations (i.e., university subscription to publisher, search engine options), we select four source databases, listed in Table 1. For the purpose of this review, we consider that the majority of the literature related to the field of information retrieval is indexed in these databases. Our first source is the ACM Digital Library, which covers a large number of computer science conferences and journals, including the proceedings of the FAccT conference (which is one of the most relevant venues for the subject), thereby making it a perfect source for this review. IEEE Xplore Digital Library is complementary to ACM Digital Library as it also indexes papers from computer science conferences and journals. The last two sources index multidisciplinary literature, which can be useful in this review as even if the main topic is information retrieval and more globally computer science, some work done in sociology or psychology might be of interest. For this review, we focus on peer-reviewed papers that assume some external quality assurance. Therefore, sources like Google Scholar[5] and arXiv[6] that allow the addition of non-reviewed papers were not considered to conduct the review. Generally, it can be assumed that the highest quality papers are covered by the selected databases and additional sources (either curated or archival) would only yield duplicates of these.

Table 1. Source databases for the review.

Source	URL
ACM Digital Library	https://dl.acm.org
IEEE Xplore Digital Library	http://ieeexplore.ieee.org
Scopus	https://scopus.com
Web of Science Core Collection	https://www.webofscience.com/wos/woscc/basic-search

2.3 Selection Criteria

Two sets of criteria are used to filter primary studies retrieved after querying the selected databases. The first set contains six inclusion criteria, while the second set is comprised of six exclusion criteria, which are detailed below.

[5]https://scholar.google.com/
[6]https://arxiv.org/search/math

Inclusion criteria.

IC1 The paper proposes a definition of one or several FATE notions with regards to IR.
IC2 The paper proposes an approach (e.g., ranking algorithm or framework) to build a trustworthy IR system.
IC3 The paper proposes a method to evaluate one or several FATE notions.
IC4 The paper presents a study that investigates the foundations of FATE in IR systems.
IC5 The paper summarises the outcomes of a workshop or a tutorial on FATE in IR systems.
IC6 The paper summarises a user study, a use case, or an audit of FATE in IR systems.

Using the different inclusion criteria, we can identify different types of primary studies. IC1 refers to Definition papers in which authors present their thoughts on the definition of FATE notions. Papers selected using IC2 are classified as Approach, while the papers describing an Evaluation method for FATE concepts are labeled with IC3. Finally, Foundation, Workshop, Tutorial are identified with IC4, IC5, and Audit, User study papers correspond to IC6.

In addition to the inclusion criteria, we also have exclusion criteria to help refine the selection of primary studies. For this review, a limited time range was defined to focus on the last 40 years, hence all papers outside of it will be removed from the pool (EC2). In case the full text of a paper is not available in the databases, we look for it on the Web, especially using scholarly search engines and paper repositories (i.e., Google Scholar, arXiv, Semantic Scholar[7]) and authors' websites. If the full text is still unavailable the paper will be excluded. IR is a vast field that includes a variety of approaches to answer information needs. However for this study, we decided to narrow the search area by focusing only on systems that receive a query expressing an information need and return a ranked list of relevant items without personalization (EC6). Moreover, papers stating that the proposed system has FATE abilities but do not give details on how and to which extent are not considered (EC5). Examples of papers excluded with regards to EC5 include tutorial descriptions [29, 87] and the approach proposed by Wu et al. [107]. In that paper, the authors "believe that this work contributes to improving ranking performance and providing more explainability for document ranking" but details on how the explanability is improved are not given.

Exclusion criteria.

EC1 The paper is not written in English.
EC2 The paper is not in the date range of January 1, 1980 to April 19, 2022.
EC3 The full-text version of the paper is not accessible.
EC4 An extended version of the paper has been published, which subsumes its contents.
EC5 There is a statement in the paper saying that the concepts of FATE are not developed.
EC6 The IR system proposed in the paper does not match with our definition of an IR system (e.g., recommender systems).

3 SYSTEMATIC REVIEW EXECUTION

After the planning phase comes the execution of a search query against the selected databases. In this section, we explain how we built a search query tailored for the topic studied in this work. Next, we provide a description of the procedure used to identify the relevant primary studies.

3.1 Search Query

In this review, there are two main aspects: FATE and IR. In order for a study to be relevant, both aspects need to be present. Consequently, each aspect corresponds to a component of a conjunctive ("AND") search query. The former aspect can be divided in four distinct terms: fairness, accountability, transparency, and ethics. Indeed,

[7]https://www.semanticscholar.org

Table 2. Search query terms and their associated supplementary terms.

Terms	Supplementary terms
fairness	fair
transparency	transparent, explainable, explanation
accountability	accountable
ethic	ethical
information retrieval	ranking algorithm, search engine

it is rare to find studies covering all aspects at once, hence the terms are combined with OR. For each notion, we identify supplementary terms such as stems and synonyms that are often mentioned when studying FATE. For example, the term "explanation" is often associated to the notion of transparency, therefore it is included in the search query. Following the definition of IR used in this review we consider ranking algorithms and search engines as supplementary terms. Table 2 summarizes the five main terms of the search query and their supplementary terms. Consequently, the final query is built as follow:

((fairness OR fair) OR (transparency OR transparent OR explainable OR explanation) OR (accountability OR accountable) OR (ethic OR ethical)) AND ("information retrieval" OR "ranking algorithm" OR "search engine")

The different databases have their own query syntax, thus the search query is customised accordingly before its execution. For each database, the search query is exclusively looking for the different terms in the title and/or the abstract of the indexed studies.

3.2 Study Selection

The search query was executed on each database on April 19, 2022 and 2,049 studies were retrieved (excluding 495 duplicates). Figure 1 illustrates the selection process via a PRISMA flow diagram [76]. This diagram includes the number of studies retrieved for each databases, as well as the number of matching studies per criterion at each step of the screening procedure.

The studies are filtered using a two-step procedure (Screening section in Figure 1) in order to keep only the ones that are considered relevant in this review. The first step consists of the analysis of the title, abstract, and keywords, if present, for all the 2,049 studies retrieved. In case the information provided by these fields indicates that the paper matches one inclusion criterion, it is kept for the second step of the filtering procedure. Conversely, the study will be removed from the pool if it matches the exclusion criterion EC2. Moreover, the selected studies are classified based on the inclusion criteria they matched. In the second step, the full text of the studies still in the pool is retrieved and examined. More specifically, we verify whether the study truly matches an inclusion criterion and validate or change its classification accordingly. Additionally, we check which study should remain in the pool based the exclusion criteria. Each study is assessed by a single researcher (the first author of the paper) who followed the described procedure to the letter. The annotated studies are made publicly available.[8]

After executing the first step of the procedure, 149 studies remained in the pool. From the original pool, 29 studies were not considered because of their publication date (i.e., matched EC2), as well as 2 studies violating IEEE publication principles. During the second step, a total of 62 studies are removed due to exclusion criteria. There are few studies such as [73, 82, 85] that match multiple inclusion criteria. Following the filtering procedure, the final pool contains a total of 75 primary studies. Table 3 shows the complete list of studies that are examined in this review.

[8]https://bit.ly/3RkQoym

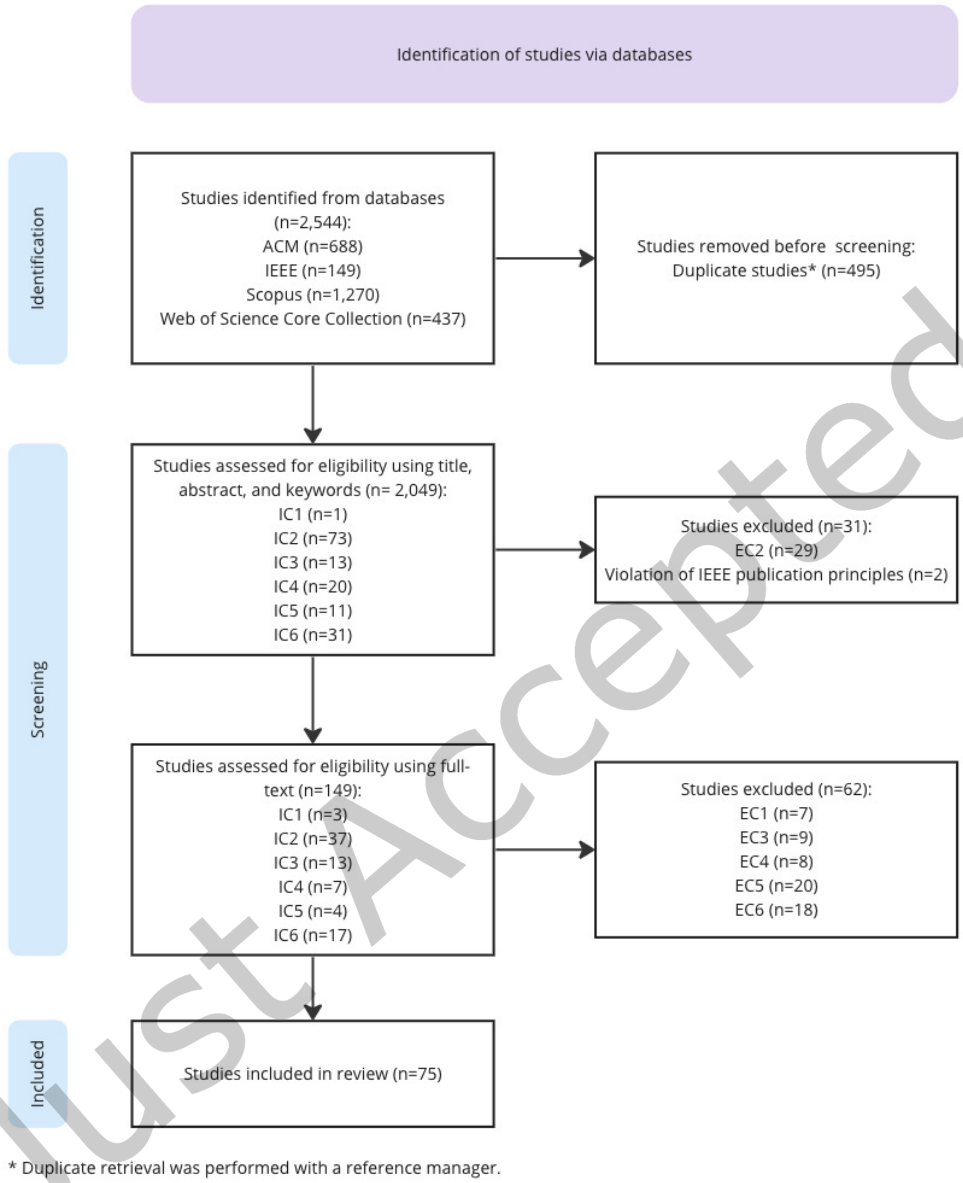

Fig. 1. PRISMA flow diagram of the selection process.

4 SYSTEMATIC REVIEW RESULTS

This section summarizes the analysis of the 75 studies selected in regards to our research questions. We first present an overview of the field and its evolution over time, then address each of the research questions in turn in Sections 4.1–4.4.

Table 3. Selected studies per criteria.

Criteria	References
Definition (IC1)	[16, 45, 82]
Approach (IC2)	[1, 4, 6, 8, 15, 17–19, 21, 23, 32, 35, 37, 42, 47, 48, 51, 57, 68, 69, 73, 74, 85, 89–91, 93, 94, 96, 99–101, 103, 110–113]
Evaluation (IC3)	[1, 4, 26, 37–39, 54, 55, 66, 79, 85, 88, 109]
Foundation (IC4)	[27, 49, 56, 62, 63, 82, 97]
Workshop, Tutorial (IC5)	[75, 80, 81, 115]
Audit, User Study (IC6)	[13, 24, 25, 33, 34, 59, 64, 70, 72, 73, 77, 84, 86, 92, 95, 106, 116]

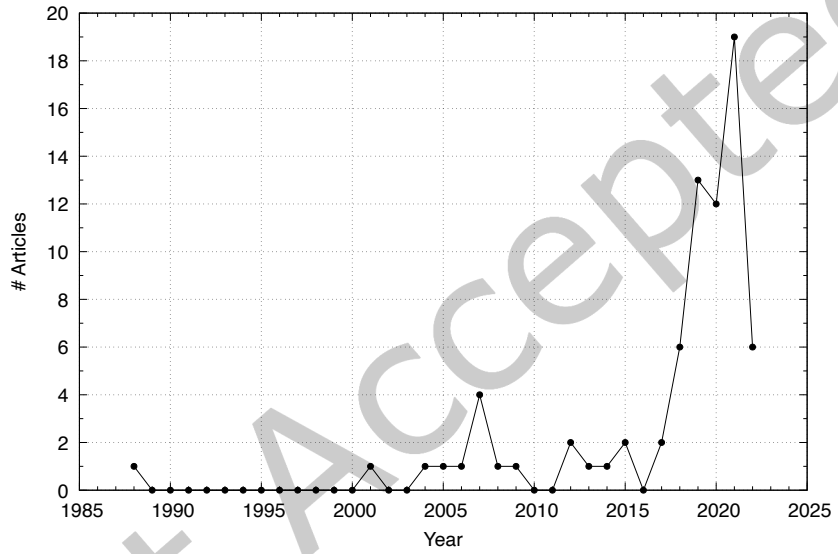

Fig. 2. Annual number of publications on FATE in IR from 1980 to April 2022.

On Figure 2, we can observe that before 2016 the publications in the field were very scarce. After that year, we see a significant rise in the number of studies, which illustrates the recent attention received by the field. We note that this peak is consistent with the development of machine learning approaches in IR, including, among others, the introduction of the Transformers model in 2017 [98]. These approaches are generally considered as black boxes that might have undesired or harmful behavior (e.g., amplifying biases) [9, 22]. Thus, it appears that the scientific community is looking into gaining insights to better understand of the inner workings of these models and to correct their undesired behavior. It is noteworthy that the different notions of FATE did not receive the same attention. Indeed, looking at the word could of the top 15 keywords extracted from the metadata of the selected studies (Figure 4), we can observe that fairness is the predominant notion ahead of transparency.

Figure 3 shows the number of studies per inclusion criteria published within 5-year periods. We can see that most of the work focuses on the development of approaches to build trustworthy IR systems, followed by evaluation methods, metrics, and audits and user studies. Indeed, looking at the last five years, we note that most

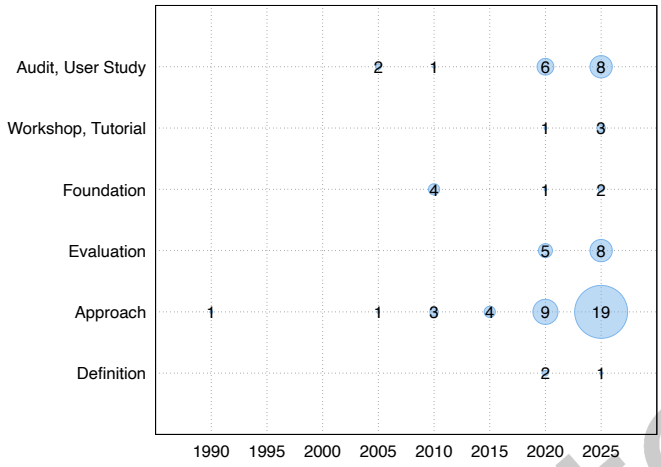

Fig. 3. Number of studies by inclusion criteria within 5-year periods.

Fig. 4. Word cloud of top 15 keywords extracted from studies metadata.

of the studies involved in the peak observed before (Figure 2) are approaches. Interestingly, there are very few studies addressing the definitions of the main concepts in the field.

4.1 RQ1: What are the definitions and characteristics of fairness, transparency, accountability, and ethics in IR?

For this first research question, we analyze Definition studies. For each study, we extract the definition and dimensions associated with FATE notions (i.e., fairness, accountability, transparency, and ethics). As a starting point, we look up the standard dictionary definition[9] of each notion, to give us an idea of what a user might expect:

Fairness: "the quality of treating people equally or in a way that is reasonable."
Accountability: "the fact of being responsible for your decisions or actions and expected to explain them when you are asked."
Transparency: "the quality of something, such as a situation or an argument, that makes it easy to understand."
Ethics: "moral principles that control or influence a person's behaviour."

The study by Hajibayova [45] is the only one that provides a definition for three notions: fairness, accountability, and transparency. Interestingly, Hajibayova [45] shifts from the traditional perspective that a system should be neutral to a human perspective and argues that it reinforces the ethical norms and standards in the information environment. Accountability is defined as the ability to justify why an information is retrievable and accessible or not, with mention to the related standards, policies, or regulations. Fairness in that case is associated with the idea that both free and copyrighted information should be retrievable and accessible, but be used responsibly. Transparency is defined as the ability of a system to provide and maintain an understandable views—for all stakeholders—of its standards, policies, and the provenance of the information.

Two other studies [16, 82] define notions in the context of ranking. Both studies state that there are several competing definitions of fairness, consistent with the elusiveness of this concept. In the machine learning and data mining communities fairness is commonly associated with the absence of discrimination (sometimes referred to as bias). Statistical discrimination due to different factors can be observed in data-driven IR systems. Castillo [16] examines the notions of fairness and transparency, while Pitoura et al. [82] focus only on the former. Castillo [16] concludes that a fair ranking has at least three characteristics:

- Ensure individual fairness (i.e., consistent treatment of similar items).
- Prevent representational harms towards a group using a proper representation of items.
- Prevent distributive/allocative harms towards a group using an adequate number of items in the ranking for each group.

Here, we find that two types of fairness emerge, i.e., individual and group fairness, depending on the granularity level of the work. Those are also reported by Pitoura et al. [82], who go further and refine the different types of fairness by proposing a more detailed taxonomy (Table 4). For example, fairness can be studied from the item producers side or not only in one ranking but on a sequence of rankings. Castillo [16] explains why having a transparent system is important by listing associated characteristics:

- Ensure alignment between system and users objectives.
- Communicate technical, specialized information in an understandable way to all stakeholders.
- Make trade-off visible.
- Support ethical compliance.
- Allow testing of claims about the system.

To summarize, we observe that the definition of *accountability* is close to the standard dictionary definition. However, for *transparency* and *fairness*, the standard definitions are found to be too vague from a technical point of view, therefore various studies have presented refinements (that are still in line with the spirit of the standard definitions). The notion of transparency relates to the ability of a system to describe its inner workings, in order

[9]https://www.oxfordlearnersdictionaries.com

Table 4. Taxonomy of fairness requirements, based on Pitoura et al. [82].

Dimensions	Sub-dimensions	Description
Level	Individual	Ensure similar treatment of similar entities (i.e., user or item).
	Group	Ensure similar treatment of entities belonging to a group. The affiliation to the group is based on the value of a protected attribute (e.g., ethnicity, gender, age).
Side	Consumer/User	Ensure that similar users or group of users receive similar ranking.
	Producer/Item	Ensure that similar items or group of items are ranked in a similar way.
Output Multiplicity	Single Output	Fairness is studied on only one output.
	Multiple Outputs	Fairness is studied on a sequence of outputs as a whole.

to communicate effectively with different stakeholders and to give them a better understanding of its output. Furthermore, it appears that there is not a one-size-fits-all definition for fairness. Indeed, studies propose different definitions depending on the application context. For example, fairness is not defined the same if we consider a single search results page (SERP) or a sequence of SERPs in the context of web search. Another example, in the context of candidate job ranking, is if we consider candidates individually or as a group (e.g, women and men). Among the studies, Pitoura et al. [82] propose the most complete exploration of fairness. Therefore, we shall use their taxonomy in the following sections to distinguish between the type of fairness studied. Finally, we notice that none of the Definition studies focuses on the notion of *ethics*.

4.2 RQ2: How to build a fair, transparent, accountable, and ethical IR system?

To answer the second research question, we analyze Approach studies. We follow the principles from grounded theory [40] to inductively derive insights from the data directly. A set of codes used to label the presented approaches is inferred from Approach studies, with the aim to capture key characteristics of the approaches. Table 5 presents the 17 codes with their description, which can be grouped into 5 main categories:

(1) **Fairness.** This category is used to characterized the notion of fairness studied. Here, we make the connection with the taxonomy proposed in [82]. With these codes, we can illustrate the diversity of definitions studied in the literature (see Figure 5).
(2) **Explanation Level.** The understanding of a system can occur at several levels, in other words the *how* and *why* of system. For example, some might be interested in the inner-workings of a search engine, while others prefer to understand why a specific output was produced.
(3) **Explanation Presentation.** The communication of the explanations should be as efficient and easily understandable as possible. Hence, explanations can take different forms: visual, textual, and structured (e.g., table).
(4) **Position in the IR process.** The introduction of FATE in an IR system can occur at different stages: before, during, or after the retrieval process.
(5) **Type.** Commonly, there are two types of approaches to solve a problem: a general one that can be applied to solve the problem in different application domains, and a specific one which is tailored for one particular domain. Additionally, a significant number of approaches consider the integration of FATE in an IR system as an optimization problem between relevance/utility and some FATE-related constraint(s).

The analysis of the characteristics approaches and the notion they study lead to the following observations. First, we notice that approaches generally focus either on fairness or transparency, through explanations. Second, a majority of the presented approaches are applied after the retrieval process, which correlates with the model

Table 5. Codes developed to characterize approaches.

Label		Description
Fairness	Group	Ensures consistent treatment of entities belonging to a group. Entities are grouped based on one or several protected attributes (e.g., race, gender, location).
	Consumer/User	Ensures that similar users or group of users receive similar ranking. For example, if political orientation is the protected attribute, every democrat should have the similar results when looking for information about gun regulations.
	Producer/Item	Ensures that similar items or groups of items are ranked in a similar way. For example, if gender is the protected attribute of a candidate, it should not impact the final ranking of candidates for a job.
	Single Output	Fairness is studied on only one output.
	Multiple Outputs	Fairness is studied on a sequence of outputs as a whole. This implies that a singular output in the sequence can exhibit unfair behaviour, while the whole sequence is considered fair.
Explanation Level	Global	Describes how the system works overall.
	Local	Describes the relationship between a specific input (i.e., query) and output (i.e., search results).
	Causal	Describes the relationship between the inner-workings of a system (i.e., the cause) and a specific output (i.e., the effect).
Explanation Presentation	Visual	Provides visual explanations (e.g., widget, graph) to understand the inner-workings and/or output of the approach.
	Textual	Provides textual explanations using natural language to understand the inner-workings and/or output of the approach.
	Structured	Provides explanation organized in a structured way (e.g., tuple, table).
Position in the IR process	Pre-process	The approach is used before the retrieval process to apply some transformations on the data (e.g., mitigate bias).
	In-process	The approach modifies the retrieval process to take into consideration at least one FATE notion.
	Post-process	The approach modifies the output of the retrieval process to take into account at least one FATE notion.
Type	Model Agnostic	The approach does not depend on the information retrieval model.
	Model Specific	The approach does depend on the information retrieval model.
	Optimization Problem	The approach solves an optimization problem between utility and at least one constraint associated to one FATE notion.

agnostic characteristic. Indeed, the majority of retrieval processes return either a list or set of top-ranked results, hence having an approach that takes this structure as input makes it easier to generalize. Next, we share insights on approaches to build a fair or transparent system.

Building a fair IR system. As reported in the previous section, fairness has multiple definitions based on different characteristics. Thus, here we analyze the proposed approaches with regards to the characteristics that are taken into consideration for the creation of a fair IR system. First, we see that all the studies focus on the group level rather than the individual one. Group fairness is particularly studied when the retrievable items are people (e.g., [18, 32, 94]). Moreover, the presented approaches examine fairness on single output except, for [96] where a sequence of outputs is examined. Thus, the aim is to have a system that is fair on average, with the possibility that some outputs are unfair. Another observation is the dominance of item-side fairness over user-side. According to Wang and Joachims [103] both sides should be studied in online platforms, because a system should treat fairly users in terms of services as well as items in terms of exposure. Hence, they propose an algorithm to optimize consumer and producer fairness at the same time. Several approaches formalize the problem of creating a fair system as an optimization problem. In other words, fairness is considered as a constraint (e.g., [17, 111]) and the goal is to find an output with the highest utility and fairness.

Building a transparent IR system. As stated before, explanations are a means to achieve transparency. There are two important characteristics for an explanation: (1) the level it describes and (2) the presentation form used for communication with stakeholders. For the former, we observe that the majority of approaches provide local explanations, which help a user to understand why the system returns a specific results. For example, Singh and Anand [89] provide an explanation for each document in the search results list to understand why the document is relevant given the terms it contains. In another example, Muramatsu and Pratt [73] show the

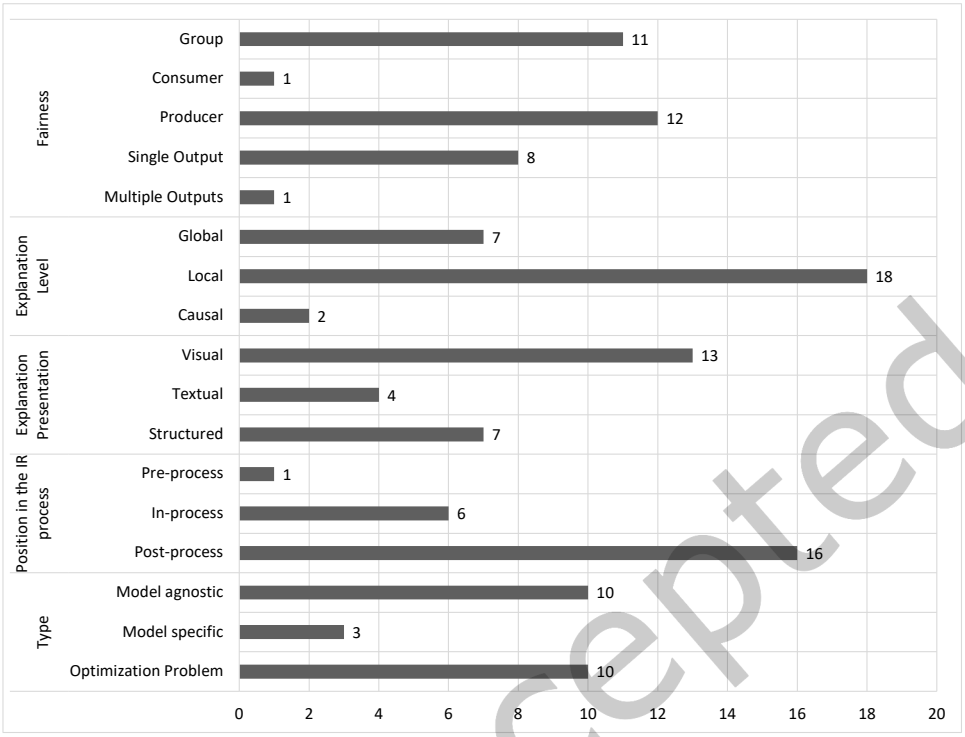

Fig. 5. Code occurrences in Approach studies.

transformations applied to the query for retrieval, which can help users understand the output produced and refine their queries. Nevertheless, we also identify two other explanation levels that provide different kinds of insights into the inner workings of systems. Belkin [8] works in a conversational setting, where the purpose of explanations is to provide insights on the abilities of the system and strategy to answer information need. An approach based on the concept of causality is presented in [68, 69], where both the inner-workings of the system and the output to a specific query are considered. The presentation of explanations often depends on the domain of application and the user interface. Indeed, in a conversational context one could favor textual explanations, like Belkin [8], while for search engines, the user interface offers more possibilities for rich visual representations. In fact, visual explanations are the most commonly used presentation form [35, 57], followed by structured representation [6, 101], and approaches that mix different presentation forms [51, 110]. Within Approach studies, only one work [100] does not use explanations to create a transparent IR system. Vilares et al. [100] state that their system is more transparent than the original one proposed in [67], because it is created with freely available resources. The advantage of using open-source resources allows every user to access and scrutinize the inner-workings of the system. This, however, is a new interpretation of transparency. Unlike the other systems proposed, the transparency is achieved "externally" through the documentation and openness of the system rather than "internally" with the production of explanations.

In summary, none of the Approach studies proposes a solution to create a system that is fair, transparent, accountable, and ethical at the same time. However, some insights on the creation of a fair or transparent system

Table 6. Evaluation datasets per domain.

Domain	Number of Evaluation	Examples of dataset
Web search	4	ClueWeb09[a], MSLR [83]
Movie	4	MovieLens [46], IMDB[b]
Justice	6	COMPAS [5]
Education	5	LSAC [105]
Publication	4	TREC 2019 Fair Ranking Track [11]
News	3	Robust04[c], TREC Common Core Track[d], TREC AQUAINT [41]
Credit	5	German Credit [28]
Retail	3	Amazon product [58]
Employment	2	Adult [28]
Other	3	args.me corpus [3], Wiki Talk Page Comments [108]

[a] https://trec.nist.gov/data/web09.html

[b] https://www.imdb.com/interfaces/

[c] https://trec.nist.gov/data/t13_robust.html

[d] https://trec.nist.gov/data/core.html

emerge. According to the research trends during the studied period, the predominant approach to building a *fair* system is a post-process, model agnostic approach that focuses on group fairness in a single output. For a *transparent* system, the same type of approach (i.e., post-process and model agnostic) is followed, which, in addition, provides a visual and local explanation.

4.3 RQ3: How are FATE notions evaluated?

Next, we look into how the different notions of FATE are evaluated. We start by checking how studies in the Approach category assessed the proposed methods with regards to the FATE notions that are being addressed. Table 6 lists the domains of evaluation with example datasets for each. Then, we analyze Evaluation studies to review metrics and methods used. Finally, we investigate Audit & User Study papers, which represent common ways of assessing a specific system with regards to FATE.

Evaluation in Approach studies. We find that the majority (74%) of approaches proposed to build a fair or transparent system come with an evaluation with regards to fairness or transparency. Furthermore, the following pattern is observed: approaches focusing on transparency tend to be evaluated with user studies, while the ones related to fairness are assessed in terms of system performance (see Figure 6). Moreover, we note that in 50% of the studies that involve performance-based evaluation, a new metric was introduced. For example, Verma and Ganguly [99] used two metrics to assess the consistency and correctness of an explanation.

Table 6 shows the various application domains that have been studied, illustrating the omnipresence of information retrieval systems in society. Various tracks at the Text REtrieval Conference (TREC), which is an annual benchmarking platform organized by the US National Institute of Standards and Technology (NIST), are one of the main source of the datasets used by Approach studies for evaluation.

Evaluation methods and metrics. Of the 13 Evaluation studies, all but two of them [1, 99] relate to the notion of fairness. Indeed, Verma and Ganguly [99] introduce a metric to evaluate the consistency of an explanation and another one for its correctness. An explanation is considered consistent if a variation of the explanation model parameters does not significantly affect it. The correctness of an explanation corresponds to higher weights given to relevant terms in the query. Abu-Rasheed et al. [1] evaluate an explainability algorithm with four metrics,

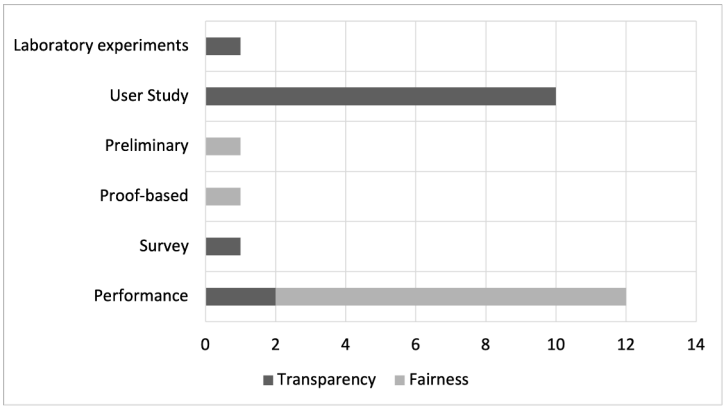

Fig. 6. Evaluation methodologies in Approach studies.

among which one is a new metric: information availability. This metric assesses the quality of an explanation based on the number of slots that are filled in an explanation template, i.e., the more slots that are filled, the better the explanation. Some studies propose a comparison of different optimization policies [38], fairness metrics [54], and diversification methods [66]. In the context of fairness as an optimization problem, Gao and Shah [38] introduce a framework to identify a solution space for a specific dataset. This space is then used to compare different optimization policies to find the optimal one. Other studies that introduce new metrics in order to quantify, estimate fairness, usually do it with a proxy. For example, Gao and Shah [39] measure bias, while Diaz et al. [26] use the concept of exposure. A different perspective is proposed by Gao et al. [37] by integrating both traditional IR metrics used for relevance assessment with fairness metrics. The fact that fairness has many definitions leads to a variety of metrics. Kuhlman et al. [54] highlight three categories: top-k, exposure, and pairwise metrics. Top-k metrics, like normalized discounted difference (rND) and normalized Discounted Ratio (rRD), are based on the idea that an item has a good outcome if it is ranked in the top-k positions. Exposure metrics focus on the attention given to an item at a specific rank position. Finally, pairwise metrics evaluate the advantage of a group compared to another one. This variety can explain why we did not identify a sort of reference metric to easily compare approaches, in a way that, e.g., mean average precision (MAP) is commonly used to compare the quality of different approaches.

Audits and user studies. The notion of accountability is generally examined through audits, as in [59, 77]. Other notions such as fairness and transparency are usually investigated through user studies; the work by Kuhlman et al. [55] is the only one that proposes an audit framework for fairness. Since there is no established standard solution for user studies, it is extremely difficult to generalize them. For example, in Audit & User Study papers, the number of participants varies between 14 [73] and 1,079 [92]. We make a similar observation for the number of questions asked in surveys/exit interviews, which ranges between 2 [95] and 21 [25]. Additionally, there are multiple options for study design, e.g., within-subject [70] or between-subject [84]. The types of questions asked from participants are also diverse: there are Likert scale questions like "By looking at the snippet, I can tell if the result is useful or not without opening the link" [70] as well as more open-ended questions such as "Try to explain, as simply as possible, how this matching rate number is calculated" [86]. Finally, the demography of the participants is usually unique to the specific study in terms of gender parity, age range, occupation, etc. Nevertheless, we notice that the participants in the selected studies are most commonly based/born in the United

States.

To conclude our findings on the evaluation of FATE in IR, we notice that a variety of metrics are available to assess *fairness*, which is likely due to the many definitions of fairness. This fact can also be linked to the observation that metrics commonly use a proxy to fairness. Interestingly, we note that only one study considers individual fairness [26]. Regarding the examination of *accountability* in a system, audits appear to be the only solution. Based on observations from a series of experiments, a discussion emerges on who should be responsible for the actions of the system. Finally, in terms of *transparency*, the common approach is to conduct user studies in order to assess users' understanding of the system. However, metrics were proposed in [1, 99], which do not evaluate transparency directly but use explanation quality as a proxy.

4.4 RQ4: What conclusions emerge from foundational and empirical studies as well as from discussions at workshops?

After investigating approaches to build trustworthy IR systems, and methods and metrics to evaluate FATE notions, we reflect on the conclusions reached in different types of studies (Approach, Evaluation, Foundational, and Workshop, Tutorial). We report the findings organized around the notion in focus.

Fairness. As stated in the previous sections, fairness is a multi-dimensional concept that does not have a one-size-fits-all definition. It is noteworthy that some tensions can be observed between different definitions [75]. Let us take individual and group fairness as an example, where it is easy to imagine that optimizing one will not optimize the other. The lack of a shared definition is a real challenge to the development of standardized solutions and evaluation metrics. Additional open challenges, such as the lack of data, are reported in [75, 82].

Previously, we observed that fairness is commonly studied through proxies such as diversity, exposure, and bias. However, an important point is that these proxies are not equivalent to fairness. In other words, having a diverse or unbiased output does not always imply that the system is fair. Pathiyan Cherumanal et al. [79] observe that some metrics capture novelty and diversity, but not on the same dimensions. Sühr et al. [92] provide another perspective on the relationship between diverse results and fairness; the authors state that a system can be diverse and improve group fairness by increasing the selection of items from underrepresented groups, but the fairness aspect entirely depends on the end user. This is illustrated by an example in the human resources domain: if the end user—a recruiter in this situation—is biased towards a gender or a race, then the overall output will not be fair.

Olteanu et al. [75] provide the following key requirements and questions that should be considered when creating a fair IR system.

(1) Create rankings based on fairer algorithms and the knowledge of possible query dependent protected attributes.
(2) What fairness constraint should be optimized to consider both the user and the item sides?
(3) Where and when should we intervene in the IR system pipeline? And, how to take limited training data with respect to sensitive attributes into consideration?
(4) Have fair user interface that may need to find a compromise between fairness and transparency.
(5) Recognize potential trade-offs between exploration and exploitation.

Transparency. We observe that transparency can have several levels and purposes (e.g., interpretability, reinforcement of trust). As a consequence, different models have been developed. At the same time, the formalization of transparency is still an open problem [75].

Ribes et al. [86] argue that the manner of presenting a detailed explanation can in some cases result in the overloading of the cognitive capacities of the user and reduce the interpretability of the system. However, if the

communication is done in a reasonable manner, the user experience is improved [25]. Another solution to make a system transparent without deteriorating the user experience is by publishing articles or blog posts about the inner-workings of the system [33, 34, 59], so that the information is available to all users who want to know more about the system. Similarly to the case of fairness discussed above, Olteanu et al. [75] provide a series of essential questions to address when creating a transparent IR system.

The vast majority of studies argue and strive for increased transparency [25, 73, 75, 89]. Unlike those, Laidlaw [56] calls for precautions, as a fully-transparent system might invite some ill-intentioned users to take advantage of it. Laidlaw [56] gives the example of search engine optimization tools, which aim to increase the traffic of a website by manipulating search engine rankings. Indeed, with total transparency, these tools would have all necessary information to dupe the ranking algorithm into considering irrelevant items as relevant.

Accountability. Studies related to the notion of accountability usually try to answer some of the following questions: (1) Who is responsible for the production of fair output? (2) Does the system follow some regulations, standards, or policies? (3) Is there an independent complaints mechanism? Laidlaw [56] argues that search engines do not belong to existing media categories, such as newspapers, hence they do not need to comply with any sort of regulations. Moreover, the author states that imposing some standards, policy, or anything making the search engine responsible, might slow down innovation due to new constraints. They identify four priorities for accountability, including the production of relevant and unbiased output and a certain degree of transparency. Lewandowski [59] takes the special case of the Google search engine and focuses on its responsibility to produce fair results. One of the points raised is that the search engine and the way the output is processed by the users are the result of human decisions, which can lead to unfairness; this is also supported by [49]. Hence, one could wonder if the responsibility falls on the end users, the engineer, or on the algorithm itself.

Ethics. Some studies look into other ethical issues not covered in the previous notions [62, 63]. MacFarlane et al. [63] investigate the question of user privacy and information bubble in IR systems. They argue for a recipient-oriented design of information systems, which would give more control to the user issuing the query, while the privacy question can be solved at the architecture level or by applying a policy. However, giving more control to users can increase their chances of ending up in an information bubble. Luyt and Lee [62] discuss at a high level the social and ethical implications of a social information retrieval system (i.e., a system to find other communities). They especially emphasize on the concepts of homophily and public sphere. Indeed, when designing such systems, it is important consider other values than relevance to break homophily.

The presence of a large gap between the beliefs, conceptualisation, and models of search engines of the end-users and the designers of search engines was shown in [13, 97]. Bilal and Zhang [13] focus on the case of teen users and report that more than half of the participants of the study were not able to explain how an output is produced by a search engine. Van Couvering [97] summarizes the output of interviews with tech professional, which led to the identification of two major schemes structuring the development of search engines: the *market schema* relates to the business side of a search engine such as costs, revenues, and user satisfaction, while the *science-technology schema* considers a search engine as a research object likely able to solve users' needs.

In this section, it is interesting to notice that a significant number of studies are driving towards fairer, more transparent, responsible, and ethical systems. However, there can be tensions within a notion (e.g., individual and group fairness), between a notion and its proxies (e.g., fairness and diversity, transparency and diversity). Additionally, some studies show a gap between the users and designers of search engines, which supports the idea that transparency can be beneficial for the user experience.

Table 7. Taxonomy of transparency requirements.

Dimensions	Sub-dimensions	Description
Degree	Full	Provide all the information necessary to make the entire system transparent.
	Partial	Provide all the information necessary to make a part of system transparent (e.g., pre-processing of the query).
Level	Global	Describe the inner-workings of the system.
	Local	Describe the relationship between a specific input (i.e., query) and output (i.e., search results).
	Causal	Describe the relationship between the inner-workings of a system (i.e., the cause) and a specific output (i.e., the effect).
Modality	User Interface (UI)	Communicate information through the UI (e.g., explanations).
	Article	Communicate information in an article, or a blog post.
	Open source	System is built on open source resources that can be scrutinized.

Table 8. Taxonomy of accountability requirements.

Dimensions	Sub-dimensions	Description
Rules	Regulation	Compliance with at least one regulation (e.g., General Data Protection Regulation (GDPR)[†]).
	Policy	Compliance with a policy.
	Standards	Compliance with some standards.
Independent complaint mechanism		Collect and process complaints independently of the system.
Responsible	User	The user is responsible of their actions based on the output.
	Algorithm	The algorithm is responsible for its own actions.
	Designer	The designer of the system is responsible for the system's actions.

[†] https://gdpr-info.eu

5 DISCUSSION

Based on what we learned through the systematic literature review, we now present our insights on trustworthy IR systems. First, we discuss about the formalization of the FATE notions through taxonomies, as well as the problems related to them. Then, in Section 5.2, we present some open challenges in the field. Finally, we conclude our discussion in Section 5.3 by acknowledging limitations of this study.

5.1 Taxonomies of Fairness, Accountability, and Transparency Requirements

Throughout this review, we observed that the different notions—fairness, accountability, transparency, and ethics—do not have clear definitions, which makes it difficult to formalize them, and consequently to develop standard solutions. *Fairness* appears to be the notion with the most number of dimensions specified. A good and detailed taxonomy for it was introduced by Pitoura et al. [82]. However, we did not find similar work for the other notions. Therefore, we develop taxonomies, by specifying dimensions and sub-dimensions, for accountability and transparency. It is worth pointing out that our taxonomies focus only on the characterization of the notions of fairness, accountability, and transparency, i.e., the requirements for an IR system to be fair, transparent, or accountable. There are additional categories we identified during the coding process, such as Position in the IR

process and Type in Table 5, which can be used to characterize ways to make IR systems more fair, accountable, and transparent, i.e., contribute to a taxonomy of methods. We believe that using these taxonomies to characterize the different notions will help to identify common trends and research gaps, in addition to relevant work when performing a comparative analysis.

For *accountability*, we propose three main dimensions, listed in Table 8. The first one describes the types of rules the system follows such as regulations and policies. The second dimension identifies if an independent complaint mechanism is available (e.g., the Information Commissioner's Office[10] in United Kingdom). Finally, the last dimension focuses on who is responsible for the system's actions—it could be the user, algorithm, or designer. In practice, the information regarding the fulfillment of these dimensions can usually be found in the "policies" and "terms and conditions" documents of IR systems. For example, search engines commonly have a privacy policy with a section related to European users and the compliance with GDPR; see, e.g., the following statement in privacy policy of Semantic Scholar:[11] "*Where personal information is transferred from the European Economic Area ("EEA") or Switzerland, we rely on appropriate safeguards such as the European Commission-approved Standard Contractual Clauses and Privacy Shield Frameworks to transfer the data and/or as otherwise authorized by applicable law.*" The responsible dimension is commonly addressed in a Liability section. However, we note here that in some cases responsibility is waived by the IR system without explicitly designating a responsible party. For example, the web policies of the National Library of Medicine,[12] which maintains PubMed a search engine for biomedical and life sciences literature, states "*For documents and software available from NLM, the U.S. Government does not warrant or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed.*" We observe that it is more difficult to find policies and terms and conditions document associated with IR systems presented at academic conferences. One of the main reason being their limited availability in time, for example, EXS search [89], MosaicSearch [74], and X-Rank [51] presented between 2018 and 2021 cannot be accessed anymore.

Similarly, we introduce a taxonomy for *transparency* in Table 7. We distinguish three main dimensions divided in refined sub-dimensions. The first one corresponds to the degree of transparency of a system, i.e., the system can be totally or partially transparent. The second sub-dimension can be used to limit the risk of malicious usage or to protect business technology. The second dimension relates to the level at which the transparency can be observed, i.e., the system, its output, or the relationship between both can be transparent. The last dimension concerns the modality used to communicate the information related to transparency. Based on the results of this review, we identify three modalities: UI, articles, and open resources. Studies presented in this survey illustrate several possible ways to implement these dimensions. For example, the Transparent Queries system [73] is partially transparent on a local level as for each query only its pre-processing is visually explained through the user interface. While, Vilares et al. [100] satisfy different sub-dimensions as they create an IR system with open source resources (e.g., Terrier[13]), hence the information to make the system fully transparent at a global level is available in the code and documentation of those resources.

The notion of *ethics* is considered an outlier in a sense that it is hardly discussed in the studies selected for this review (less than 10% of the studies focus on *ethics*). In our opinion, the fact that it is not discussed specific to IR suggests two possibilities. Either the question of ethics is not specific to IR but applies more generally to AI. Indeed, there are surveys of ethical guidelines for AI [44, 50] as well as work on the ethical design of AI systems [14]. It could also be that ethics has IR-specific aspects, but the current focus is on the other notions of FATE, which are easier to characterize but challenging nevertheless. Consequently, it might take some more time before ethics in IR receives sufficient attention. Either way, existing studies suggest that the various ethical

[10]https://ico.org.uk
[11]https://allenai.org/privacy-policy
[12]https://www.nlm.nih.gov/web_policies.html
[13]http://terrier.org

considerations, such as confidentiality and safety [75], are each complex enough to deserve their own taxonomy to be developed. This, however, is beyond the scope of this work.

5.2 Open Challenges

From our observations, it appears that there are still many open challenges in the field. Few of them are listed below. On a high-level, we identify two main direction of work. The first one is the definition and formalization of the problems as discussed in the previous sections. Secondly, an important challenge is the evaluation of each notion and, more particularly, the development of standardized evaluation protocols and metrics.

- **Individual fairness**. In this review, only one of the selected studies [26] reports on individual fairness, while there is a lot of work on group fairness. As stated before, individual fairness aims to ensure a consistent treatment of similar items or users. Our review did not yield a clear explanation as to why individual fairness did not receive the same attention as group fairness. This is all the more interesting because according to Biega et al. [12] these two are related; in fact, the authors state "when equity of attention is achieved for individuals, it will also be achieved at the group level." Other work [36, 114] present some conflicts that can occur between different kinds fairness. Hence, we believe that a better treatment of individual fairness in the literature would contribute to study potential trade-offs with group fairness and provide better guidelines for the development of future IR systems.
- **Regulations**. The question about the regulation of IR system remains open despite the widespread deployment of IR systems for several decades now. Countries tend to have different standards, while IR systems are used internationally, thus one can wonder if a general agreement can be reached. An example is the policies on data privacy, in Europe there is the GDPR while another one is applied in California (California Consumer Privacy Act). However, the operationalization of some aspects of these regulations remains challenging and unclear to date [20]. For instance, there are obstacles related to the operationalization of data minimization (Article 5(1)(c) in GDPR), especially the lack of guidelines with a practical definition and information on when and how data minimization should be applied, as pointed out by Biega [10] in their talk given at the 13th European Summer School in Information Retrieval (ESSIR '22). Therefore, we believe that interdisciplinary communication is essential to develop regulations that are socially accepted and technically realistic.
- **Ethics**. In this review, we observe that the notion of ethics is sparsely studied compared to other notions. A specific reason for this could not be deducted from our review. However, it is worth pointing out that ethics is a broad term that comprises several sub-notions such as privacy and safety, thus, its formalization in the context of information retrieval is non-trivial. Therefore, we believe that a study of each sub-notion under ethics, as a first step, would help the research community to have a better understanding of what is at stake when talking about ethics. Then, the findings should enable to make progress in the long term.
- **Benchmarks**. The evaluation of IR systems against public benchmarks is more than a common practice, it is the very trademark of the field [102]. In the last few years, some benchmarks addressing the notion of fairness have been initiated. For example, the TREC Fair Ranking Track[14] has started in 2019 and is running annually since. In 2019, the track targeted group fairness on the producer side and multiple outputs [11]. The tasks addressed by the track evolve every year, therefore the targeted (sub-)dimensions can change accordingly. Most recently, the NTCIR Fair Web Task[15] has been introduced in 2022. This web search task focuses on item group fairness in a single output. To the best of our knowledge, there are no benchmarks for the other notions (i.e., accountability, transparency, and ethics). The creation of a benchmark for accountability is especially challenging as evaluation metrics do not exist for it yet. As

[14]https://fair-trec.github.io/index.html
[15]http://sakailab.com/fairweb1/

stated before, the case of ethics is even more challenging, as the various ethical considerations would need to be formally characterized first. Overall, the development benchmarks corresponding to each FATE notion would be a critical enabler of progress. Moreover, such benchmarks would allow for standardized comparisons between information retrieval systems.

As IR systems are more and more used to automate tasks that have a direct impact on people's lives, such as candidate ranking and news retrieval. One question we ask ourselves after this literature review is: Is it possible to combine all the work on the different notions to create a system that is fair, transparent, accountable, and ethical all at the same time? If not, what are the trade-offs and who should be deciding which notion to favor over the others? For example when looking for a job, one does not want to be ranked lower in the pool of competent candidates because of protected attributes (e.g., gender, nationality), and the recruiter might want to know how the pool of candidates is created to identify potential biases and report them to the correct entity (i.e., the designer of the system). However, this is only valid when protected attributes are known via the candidate's application; indeed, if such attributes are not specified or deliberately withhold, the information retrieval system cannot be blamed. Tension could also arise between transparency and ethics, more particularly confidentiality. Indeed, in order to be transparent, the system needs to reveal information; however, it is important not to leak confidential information. Take the example of information retrieval in the medical domain, where the explanation of a ranking may contain the patient's personal data that should not be shared with everyone.

5.3 Limitations

The aim of this systematic review was to provide a comprehensive overview of what has been done in the field of FATE in IR since 1980 based on a large selection of studies. It could be argued that the annotation of the studies is subjective and biased. Indeed, it is possible that the study selection and classification would have been different if done by another researcher. However, the candidate set of papers and their annotations are made publicly available to support reproducibility.

It is worth noting that this review does not cover all available studies in the field caused by the methodical procedure of gathering research from a limited number of sources. An example is the work by Zimmer [117] that is not indexed in the selected sources. However, we chose the source databases based on their domain and the number of papers indexed in order to provide a high level of coverage of relevant studies. Despite this, the methodical approach for this review offers the possibility to be extended in future works using other sources and/or time range. It is also possible that the search terms do not appear in the title or abstract of relevant studies. An example is the work by Biega et al. [12] that does not include the terms "ranking algorithm", "information retrieval", or "search engine" in its title or abstract.

This work is based on an arguably narrow definition of IR systems: rank items in response to a textual query. A broad definition of a modern IR system is to provide the "right information, in the right way, at the right time." Still, our findings show that even with a narrow definition the subject is complex and many challenges remain. Therefore, the generalization to an extended definition of IR systems is left for future work. This especially includes contemporary IR systems relying on users personal activities to enhance their performance and deliver tailored results.

6 CONCLUSION

The field of challenges of fairness, accountability, transparency, and ethics in information retrieval has attracted a lot of attention lately. This can be explained by the omnipresence of IR systems in the society as well as the wish to automatize data-intensive tasks in order to answer specific requests (e.g., retrieving news, ranking job candidates). To be well accepted, these systems should comply to ethical norms such as being fair, transparent, and respect users' privacy. This is not always easy to achieve, especially with systems based on machine learning

models, which are often considered as black boxes. Therefore, a systematic literature review was performed to give an overview of the field. More specifically, this review analysed 75 studies focusing on the following points:

(1) the definitions of fairness, accountability, transparency, and ethics;
(2) the solutions proposed to build a fair, accountable, transparent, and ethical IR system;
(3) the evaluation methodologies used to assess the solutions proposed;
(4) the conclusions emerging from foundational and empirical studies as well as from discussions at workshops.

From this review, we learned that the definitions of the different notions are complex due to their multi-dimensional nature, which makes it challenging to establish shared standards. Following the principles of grounded theory, we found that there is no work combining all of the notions together. However, we established that the proposed approaches can intervene at different stages of the information retrieval process and can be model specific or not. The trends suggest preferred approaches to build a fair or transparent IR system. These are post-process and model agnostic approaches that study group fairness in a single ranking or provide visual and local explanations respectively. With regards to evaluation, there are a variety of metrics to automatically assess fairness, this is likely due to the different possible definitions of fairness. While the evaluation of accountability and transparency is commonly performed with different methodologies such as audits and user studies. Finally, the conclusions reached by different studies provide some perspectives regarding the creation of ethical IR systems. More particularly, some tensions between different notions and proxies associated to them were found. In addition, a gap between beliefs of the designers, engineers and users of IR systems was identified.

A taxonomy of fairness was previously proposed [82], however similar taxonomies for the other notions of FATE were not found in this review. Therefore, we contributed to filling this gap by introducing taxonomies for transparency and accountability based on (sub-)dimensions identified during the review. The development of a taxonomy for ethics is left for future work, due to the low representation of ethics in the selected studies. We hope that this is a step towards a better understanding of these notions, formalization of the problems encountered in the field, and the development of standards. We strongly believe that collaboration between researchers from different fields such as legal studies, sociology, and information retrieval to refine and extend these dimensions would be beneficial.

Based on the open challenges identified from the literature review, we highlight a few points related to FATE that should be considered for the development and deployment of future IR systems. First, it is important to clearly define the notion studied; this is especially true for fairness as many definitions exist, and certain definitions may not align or be compatible with others. Furthermore, the different trade-offs considered between the notions and/or the performance should be made clear, so the end-users and experts can better understand the system's results and behavior. Second, the development of new evaluation methodologies and metrics for the different notions should help to prevent harmful behavior before the deployment of the IR system. Third, the regulations of IR systems with regards to FATE notions could be review by a team a multi-disciplinary experts in order to be more precise, in accordance with society's beliefs, and technically realistic.

ACKNOWLEDGMENTS

This research was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, project number 309834).

REFERENCES

[1] Hasan Abu-Rasheed, Christian Weber, Johannes Zenkert, Mareike Dornhöfer, and Madjid Fathi. 2022. Transferrable Framework Based on Knowledge Graphs for Generating Explainable Results in Domain-Specific, Intelligent Information Retrieval. *Informatics* 9, 1 (2022).

[2] Deepak Agarwal and Maxim Gurevich. 2012. Fast Top-k Retrieval for Model Based Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. 483–492.

[3] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me Corpus. In *KI 2019: Advances in Artificial Intelligence (KI '19)*. 48–59.
[4] Abdulaziz AlQatan, Leif Azzopardi, and Yashar Moshfeghi. 2020. Analyzing the Influence of Bigrams on Retrieval Bias and Effectiveness. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR '20)*. 157–160.
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2022-08-12.
[6] Zhifeng Bao, Yong Zeng, Tok Wang Ling, Dongxiang Zhang, Guoliang Li, and H. V. Jagadish. 2015. A General Framework to Resolve the MisMatch Problem in XML Keyword Search. *The VLDB Journal* 24, 4 (2015), 493–518.
[7] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact Essay. *California Law Review* 104, 3 (2016), 671–732.
[8] N. J. Belkin. 1988. On the Nature and Function of Explanation in Intelligent Information Retrieval. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '88)*. 135–145.
[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. 610–623.
[10] Asia Biega. 2022. Responsible Design of Information Access Systems. http://essir2022.org/slides/asia-biega.pdf Accessed: 2022-10-13.
[11] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2019. Overview of the TREC 2019 Fair Ranking Track. In *The Twenty-Eighth Text REtrieval Conference Proceedings (TREC '19)*.
[12] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 405–414.
[13] Dania Bilal and Yan Zhang. 2021. Teens' Conceptual Understanding of Web Search Engines: The Case of Google Search Engine Result Pages (SERPs). In *Human-Computer Interaction. Design and User Experience Case Studies (HCII '21)*. 253–270.
[14] Joanna Bryson and Alan Winfield. 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* 50, 5 (2017), 116–119.
[15] Ian Burke, Robin Burke, and Goran Kuljanin. 2021. Fair Candidate Ranking with Spatial Partitioning: Lessons from the SIOP ML Competition. In *Proceedings of the First Workshop on Recommender Systems for Human Resources co-located with the 15th ACM Conference on Recommender Systems (RecSysHR '21)*.
[16] Carlos Castillo. 2019. Fairness and Transparency in Ranking. *ACM SIGIR Forum* 52, 2 (2019), 64–71.
[17] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP '18)*. 28:1–28:15.
[18] Mattia Cerrato, Marius Köppel, Alexander Segner, Roberto Esposito, and Stefan Kramer. 2020. Fair Pairwise Learning to Rank. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA '20)*. 729–738.
[19] Adrian-Gabriel Chifu, Josiane Mothe, and Md Zia Ullah. 2020. Fair Exposure of Documents in Information Retrieval: a Community Detection Approach. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE '20)*.
[20] Mark Coeckelbergh. 2019. Artificial Intelligence: Some Ethical Issues and Regulatory Challenges. *Technology and Regulation* 2019 (2019), 31–34.
[21] Gautam Das, Vagelis Hristidis, Nishant Kapoor, and S. Sudarshan. 2006. Ordering the Attributes of Query Results. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*. 395–406.
[22] Chris DeBrusk. 2018. The Risk of Machine Learning Bias (And How to Prevent It). https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/ Accessed: 2022-08-22.
[23] Aditya Dey, Chandan Radhakrishna, Nishitha Nancy Lima, Suraj Shashidhar, Sayantan Polley, Marcus Thiel, and Andreas Nürnberger. 2021. Evaluating Reliability in Explainable Search. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS '21)*. 1–4.
[24] Giorgio Maria Di Nunzio, Alessandro Fabris, Gianmaria Silvello, and Gian Antonio Susto. 2021. Incentives for Item Duplication Under Fair Ranking Policies. In *Advances in Bias and Fairness in Information Retrieval (BIAS '21)*. 64–77.
[25] Cecilia di Sciascio, Eduardo Veas, Jordan Barria-Pineda, and Colleen Culley. 2020. Understanding the Effects of Control and Transparency in Searching as Learning. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. 498–509.
[26] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. 275–284.
[27] Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. 2015. Navigating Controversy as a Complex Search Task. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks co-located with the 37th European Conference on Information Retrieval (SCST '15)*.
[28] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[29] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Retrieval and Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 1403–1404.

[30] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.
[31] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. 49–56.
[32] Yunhe Feng, Daniel Saelid, Ke Li, Ruoyuan Gao, and Chirag Shah. 2021. Towards Fairness-Aware Ranking by Defining Latent Groups Using Inferred Features. In *International Workshop on Algorithmic Bias in Search and Recommendation (BIAS '21)*. 1–8.
[33] Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, James Ostell, and Zhiyong Lu. 2018. Best Match: New relevance search for PubMed. *PLOS Biology* 16, 8 (2018), 1–12.
[34] Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How User Intelligence is Improving PubMed. *Nature Biotechnology* 36, 10 (2018), 937–945.
[35] Björn Forcher, Thomas Roth-Berghofer, Stefan Agne, and Andreas Dengel. 2014. Intuitive Justifications of Medical Semantic Search Results. *Engineering Applications of Artificial Intelligence* 30 (2014), 1–17.
[36] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Commun. ACM* 64, 4 (2021), 136–143.
[37] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR: Fairness-Aware Information Retrieval Evaluation. *Journal of the Association for Information Science and Technology* (2022), 1–13.
[38] Ruoyuan Gao and Chirag Shah. 2019. How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. 229–236.
[39] Ruoyuan Gao and Chirag Shah. 2020. Toward Creating a Fairer Ranking in Search Engine Results. *Information Processing & Management* 57, 1 (2020).
[40] Barney G Glaser. 1992. *Basics of Grounded Theory Analysis: Emergence Vs. Forcing*. Sociology Press.
[41] David Graff. 2002. The AQUAINT Corpus of English News Text. https://catalog.ldc.upenn.edu/LDC2002T31
[42] Maurice Grant, Adeesha Ekanayake, and Douglas Turnbull. 2013. Meuse: Recommending Internet Radio Stations. In *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR '13)*. 281–286.
[43] Michael Gusenbauer and Neal R. Haddaway. 2020. Which Academic Search Systems are Suitable for Systematic Reviews or Meta-analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed, and 26 Other Resources. *Research Synthesis Methods* 11, 2 (2020), 181–217.
[44] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30, 1 (2020), 99–120.
[45] Lala Hajibayova. 2019. Guardians of the Knowledge: Relevant, Irrelevant, or Algorithmic? *Information Research* 24, 4 (2019).
[46] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015), 1–19.
[47] Sam Hepenstal, Leishi Zhang, Neesha Kodagoda, and B. L. William Wong. 2020. What are you Thinking? Explaining Conversation Agent Responses for Criminal Investigations. In *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with 25th International Conference on Intelligent User Interfaces (ExSS-ATEC '20)*.
[48] Sam Hepenstal, Leishi Zhang, Neesha Kodagoda, and B. L. William Wong. 2021. *A Granular Computing Approach to Provide Transparency of Intelligent Systems for Criminal Investigations*. 333–367.
[49] Lucas D. Introna. 2007. Maintaining the Reversibility of Foldings: Making the Ethics (Politics) of Information Technology Visible. *Ethics and Information Technology* 9, 1 (2007), 11–25.
[50] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.
[51] Jian Kang, Scott Freitas, Haichao Yu, Yinglong Xia, Nan Cao, and Hanghang Tong. 2018. X-Rank: Explainable Ranking in Complex Multi-Layered Networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1959–1962.
[52] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. 3819–3828.
[53] Barbara Ann Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University and Durham University Joint Report.
[54] Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. 674–682.
[55] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference (WWW '19)*. 2936–2942.
[56] Emily B. Laidlaw. 2008. Private Power, Public Interest: An Examination of Search Engine Accountability. *International Journal of Law and Information Technology* 17, 1 (2008), 113–145.

[57] Lars Langer and Erik Frøkjær. 2008. Improving Web Search Transparency by Using a Venn Diagram Interface. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges (NordiCHI '08)*. 249–256.
[58] Jure Leskovec and Rok Sosič. 2016. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology* 8, 1 (2016).
[59] Dirk Lewandowski. 2017. Is Google Responsible for Providing Fair and Unbiased Results? In *The Responsibilities of Online Service Providers*. 61–77.
[60] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.
[61] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundation and Trends in Information Retrieval* 3, 3 (2009), 225–331.
[62] Brendan Luyt and Chu Keong Lee. 2008. The Ethics of Social Information Retrieval. In *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively*. 179–188.
[63] Andrew MacFarlane, Sondness Missaoui, Stephann Makri, and Marisela Gutierrez Lopez. 2022. Sender vs. Recipient-Orientated Information Systems Revisited. *Journal of Documentation* 78, 2 (2022), 485–509.
[64] Marcel Machill, Christoph Neuberger, Wolfgang Schweiger, and Werner Wirth. 2004. Navigating the Internet: A Study of German-Language Search Engines. *European Journal of Communication* 19, 3 (2004), 321–347.
[65] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
[66] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2022. Search Results Diversification for Effective Fair Ranking in Academic Search. *Information Retrieval Journal* 25, 1 (2022), 1–26.
[67] Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7, 1 (2004), 73–97.
[68] Massimo Melucci. 2020. Some Reflections on the Use of Structural Equation Modeling for Investigating the Causal Relationships that Affect Search Engine Results. In *Proceedings of the First Workshop on Bridging the Gap between Information Science, Information Retrieval and Data Science co-located with 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (BIRDS '20)*. 100–109.
[69] Massimo Melucci and Adriano Paggiaro. 2019. Evaluation of Information Retrieval Systems Using Structural Equation Modeling. *Computer Science Review* 31 (2019), 1–18.
[70] Siyu Mi and Jiepu Jiang. 2019. Understanding the Interpretability of Search Result Summaries. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 989–992.
[71] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundation and Trends in Information Retrieval* 13, 1 (2018), 1–126.
[72] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring Search Engine Bias. *Information Processing & Management* 41, 5 (2005), 1193–1205.
[73] Jack Muramatsu and Wanda Pratt. 2001. Transparent Queries: Investigation Users' Mental Models of Search Engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. 217–224.
[74] Yuta Nemoto and Vitaly Klyuev. 2021. Tool to Retrieve Less-Filtered Information from the Internet. *Information* 12, 2 (2021).
[75] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D. Ekstrand, 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. 2021. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. *ACM SIGIR Forum* 53, 2 (2021), 20–43.
[76] Matthew J. Page, Joanne E. McKenzie, , Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews* 10, 89 (2021).
[77] Orestis Papakyriakopoulos and Arwa M. Mboya. 2022. Beyond Algorithmic Bias: A Socio-Computational Interrogation of the Google Search by Image Algorithm. *Social Science Computer Review* (2022).
[78] The European Parliament and the Council of the European Union. 2021. The Artificial Intelligence Act. https://artificialintelligenceact.eu Accessed: 2022-09-01.
[79] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2021. Evaluating Fairness in Argument Retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. 3363–3367.

[80] Evaggelia Pitoura, Georgia Koutrika, and Kostas Stefanidis. 2020. Fairness in Rankings and Recommenders. In *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT '20)*. 651–654.
[81] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness-aware Methods in Rankings and Recommenders. In *2021 22nd IEEE International Conference on Mobile Data Management (MDM '21)*. 1–4.
[82] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *The VLDB Journal* 31, 3 (2021), 431–458.
[83] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. arXiv:1306.2597 [cs.IR]
[84] Jerome Ramos and Carsten Eickhoff. 2020. Search Result Explanations Improve Efficiency and Trust. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1597–1600.
[85] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 306–316.
[86] Delphine Ribes, Nicolas Henchoz, Hélène Portier, Lara Defayes, Thanh-Trung Phan, Daniel Gatica-Perez, and Andreas Sonderegger. 2021. Trust Indicators and Explainable AI: A Study on User Perceptions. In *Human-Computer Interaction – INTERACT 2021 (INTERACT '21)*. 662–671.
[87] Rishiraj Saha Roy and Avishek Anand. 2020. Question Answering over Curated and Open Web Sources. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 2432–2435.
[88] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. 553–562.
[89] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search Using Local Model Agnostic Interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. 770–773.
[90] Christian Stab, Kawa Nazemi, Matthias Breyer, Dirk Burkhardt, and Jörn Kohlhammer. 2012. Semantics Visualization for Fostering Search Result Comprehension. In *The Semantic Web: Research and Applications (ESWC '12)*. 633–646.
[91] Mohameth-François Sy, Sylvie Ranwez, Jacky Montmain, Armelle Regnault, Michel Crampes, and Vincent Ranwez. 2012. User Centered and Ontology Based Information Retrieval System for Life Sciences. *BMC Bioinformatics* 13, 1 (2012).
[92] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. 989–999.
[93] Shayan A. Tabrizi and Azadeh Shakery. 2019. Perspective-based Search: A New Paradigm for Bursting the Information Bubble. *FACETS* 4, 1 (2019), 350–388.
[94] Saedeh Tahery, Seyyede Zahra Aftabi, and Saeed Farzi. 2021. A GA-based Algorithm Meets the Fair Ranking Problem. *Information Processing & Management* 58, 6 (2021).
[95] Paul Thomas, Bodo Billerbeck, Nick Craswell, and Ryen W. White. 2019. Investigating Searchers' Mental Models to Inform Search Explanations. *ACM Transactions on Information Systems* 38, 1 (2019), 1–25.
[96] Thibaut Thonet and Jean-Michel Renders. 2020. Multi-grouping Robust Fair Ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 2077–2080.
[97] Elizabeth Van Couvering. 2007. Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality. *Journal of Computer-Mediated Communication* 12, 3 (2007), 866–887.
[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS '17)*. 5998–6008.
[99] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 1281–1284.
[100] Jesús Vilares, Michael P Oakes, and Manuel Vilares. 2007. A Knowledge-light Approach to Query Translation in Cross-language Information Retrieval. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP '07)*. 624–630.
[101] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '21)*. 13–22.
[102] Ellen Voorhees and Donna Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.
[103] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '21)*. 23–41.
[104] Qi Wang, Constantinos Dimopoulos, and Torsten Suel. 2016. Fast First-Phase Candidate Generation for Cascading Rankers. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 295–304.
[105] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.
[106] Fons Wijnhoven and Jeanna van Haren. 2021. Search Engine Gender Bias. *Frontiers in Big Data* 4 (2021).

[107] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Jingtao Zhan, Yukun Zheng, Min Zhang, and Shaoping Ma. 2020. Leveraging Passage-Level Cumulative Gain for Document Ranking. In *Proceedings of The Web Conference 2020 (WWW '20)*. 2421–2431.
[108] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 1391–1399.
[109] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI '19)*. 6035–6042.
[110] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A Nutritional Label for Rankings. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. 1773–1776.
[111] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. 1569–1578.
[112] Meike Zehlike, Tom Sühr, Carlos Castillo, and Ivan Kitanovski. 2020. FairSearch: A Tool For Fairness in Ranked Search Results. In *Companion Proceedings of the Web Conference 2020 (WWW '20)*. 172–175.
[113] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with Multiple Protected Groups. *Information Processing & Management* 59, 1 (2022).
[114] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *Comput. Surveys* 55, 6 (2022), 1–36.
[115] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2019. Report on EARS'18: 1st International Workshop on ExplainAble Recommendation and Search. *ACM SIGIR Forum* 52, 2 (2019), 125–131.
[116] Futao Zhao, Zhong Yao, Biao Xu, and Pengfei Tang. 2018. Exploring Fairness and Accuracy of Retrieval Models. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD '18)*. 986–992.
[117] Michael Zimmer. 2010. *Web Search Studies: Multidisciplinary Perspectives on Web Search Engines*. 507–521.