Involve Humans in Algorithmic Fairness Issue: A Systematic Review

Dan Wu[1,2](✉) and Jing Liu[1]

[1] School of Information Management, Wuhan University, Wuhan 430072, China
woodan@whu.edu.cn

[2] Center of Human-Computer Interaction and User Behavior,
Wuhan University, Wuhan 430072, China

Abstract. With the increasing penetration of technology into society, algorithms are more widely used in people's lives. The intentional or unintentional bias brought about by algorithms may affect people's lives and even affect the destiny of certain groups of people, which raises concerns about algorithmic fairness. We aim to systematically explore the current research of human-centered algorithmic fairness (HAF) research, understand how to involve human in algorithmic fairness issue and how to promote algorithmic fairness from the human perspective. This review followed the procedure of systematic review, identifying 417 articles of algorithmic fairness ranging from the years 2000 to 2020 from 5 target databases. Application of the exclusion criteria led to 26 included articles, which are highly related to human-centered algorithmic fairness. We classified these works into 4 categories based on their topics and concluded the research scheme. Methodological conclusions are presented from novel dimensions. Besides, we also summarized 3 patterns of human-centered algorithmic fairness. Research gaps and suggestions for future research also be discussed in this review based on the findings of current research.

Keywords: Algorithmic fairness · Human-centered · Systematic review · Algorithmic bias

1 Introduction

Fueled by the ever-growing development of artificial intelligence (AI) technology, algorithms are often embedded in a wide variety of systems and are increasingly employed to make consequential decisions for human subjects. However, the algorithmic bias/discrimination issues have aroused a lot of concern recently due to their potential impact on human lives.[1] believe that algorithm might be inherently biased since it learns and preserves historical biases. Algorithmic bias was found in various scenarios. For instance, in the field of criminal justice, some studies showed the algorithm used by the criminal justice system (COMPASS) has a preference for white people since it falsely predicted future criminality among African-American [2]. Another example is advertisement. It was shown that Google's ad-targeting algorithm proposed higher-paying job advertisements for men than for women. [3].

M. Smits (Ed.): iConference 2022, LNCS 13192, pp. 161–176, 2022.
https://doi.org/10.1007/978-3-030-96957-8_15

Luckily, a lot of scholars have explored how to develop fairer algorithms from a technical perspective. They proposed different definitions of algorithmic fairness, including disparate impact [4], demographic parity [5], equalized odds [6], equal opportunity [6], and individual fairness [7]. Each definition has its own formula and metrics that can help evaluate whether the algorithm is fair. But it's impossible to satisfy multiple notions of algorithmic fairness simultaneously [8, 9]. Indeed, many studies support the existence of trade-off between fairness and accuracy from both theoretical and empirical perspectives [10]. Thus, how to achieve a model that allows for higher fairness without significantly compromising the accuracy or other alternative notion of utility still needs to be explored.

However, in addition to the technical perspective, we believe that algorithmic fairness research should also be conducted from a human perspective. The reasons are: (i) Algorithms serve human beings, and some algorithmic decisions may even have a lifelong impact on humans. Therefore, human feelings and cognitions of algorithmic fairness should also be considered. (ii) Compared to technical data training, human can evaluate fairness and utility more directly based on their own needs, which may provide a different solution for developing algorithms that balance fairness and utility. (iii)In fact, it has become the consensus of many scholars that algorithmic fairness is not only a technical problem but should be regarded as a sociotechnical issue [11].

Motivated by this, we conducted this systematic review to explore how to involve humans in algorithmic fairness issue. To be more specific, what can be done, what methods could be used, and how to accompanist technical research.

To the best of our knowledge, there is no systematic literature review to date investigating algorithmic fairness from the human perspective. It's helpful to emphasize the importance of human on algorithmic fairness and provide some research clues for scholars interested in this topic.

2 Methodology

We put forward and define human-centered algorithmic fairness (HAF) as "exploring algorithmic fairness from the human point of view." To be more specific, HAF regard human as an important part of algorithm development and emphasizing the impact of human (including users, developers, practitioners, etc.) on algorithmic fairness.

A systematic review focuses on the detailed research question and aims to provide evidence of a subject or research area [12]. To conduct this study, we follow the four-element framework put forward by [13], which including a) set up the review question(s); b) mapping and scoping the review space; c) reviewing, evaluating, and synthesizing extant research base, and d) devising systematic empirical evidence drawn from the reviewed articles.

We proposed our detailed research questions based on the research goal put forward in the introduction part and established our research protocol. We designed the research questions as follows:

RQ1: What are the research topics and the research scheme of HAF research?
RQ2: What kinds of research methods are appropriate for HAF research?
RQ3: How to involve humans in the whole process of algorithm development?

2.1 Search Terms and Retrieval Strategy

Although we focus on articles that explore algorithmic fairness, the occurrence of algorithmic bias has led to discussion among society on the topic of algorithm fairness. We believe that there is a twin relationship between algorithmic bias and algorithmic fairness. Thus, we also include algorithmic bias as search terms. We expanded the search terms based on "algorithmic fairness" and "algorithmic bias" and developed a Boolean search string as follows: (algorithm* AND (fairness OR equity OR bias OR discrimination)). We adopt this search string in the article title, abstract, and keywords. We limited the time to "2000–2020" and the language to "English".

The strategy above was applied to the following five databases: WOS, ACM digital library, APA, Wiley online library, and Elsevier ScienceDirect. The search for databases was modified to fit the specific settings for each database due to the different characteristics of databases, such as adjustment of thesaurus keywords and limitation settings.

The initial search was conducted on November 23, 2020. All data were imported into Endnotes for management, and duplications were then removed.

2.2 Selection Strategy

Two rounds of screening were conducted to select relevant articles.

The first round of screening is the title and abstract screening. In this part, we aimed to filter out the articles related to the topic of algorithm fairness. Based on this, we used the following four selection criteria for article exclusion in this phase:

- Non-English articles.
- Domain irrelevant: Articles unrelated to algorithms.
- Topic irrelevant: Articles unrelated to algorithm fairness.
- Low topic relevance: Articles briefly mention algorithm fairness, but algorithm fairness is not the main content of its research.

This round of screening was conducted by a trained reviewer according to the criteria. The screening is performed twice and at intervals of one month. If there is a difference between the two times, a second person is brought into judge, ultimately reaching a final consensus of the coding and article inclusion. We regard articles left after the first round of screening as potentially eligible.

The second round of screening is full-text screening. In order to ensure articles included are fit to our research goals and RQs, we set up another 2 criteria to limit the inclusion of articles:

- Not focuses on human perspective: As we mentioned at the beginning of this section, this study focuses on Human-centered perspective, that is, emphasizing the impact of human (including users, developers, practitioners, etc.) on algorithmic fairness or using methods of user study. Articles that do not meet this qualification will be excluded.

- Articles with low quality. We evaluate the quality of the articles based on three aspects: (i) whether the research is complete, (ii) whether the method is appropriate, and (iii) whether the conclusion is representative.

2.3 Data Extraction Strategy

A top-down code of articles' content was developed for data extraction. We captured general information (author, title, publication year) and methodology components (methods, sample size, and sample group background) of included articles.

We also conducted a content analysis of the included articles in terms of research topics, research methods, and research patterns by browsing through the full text.

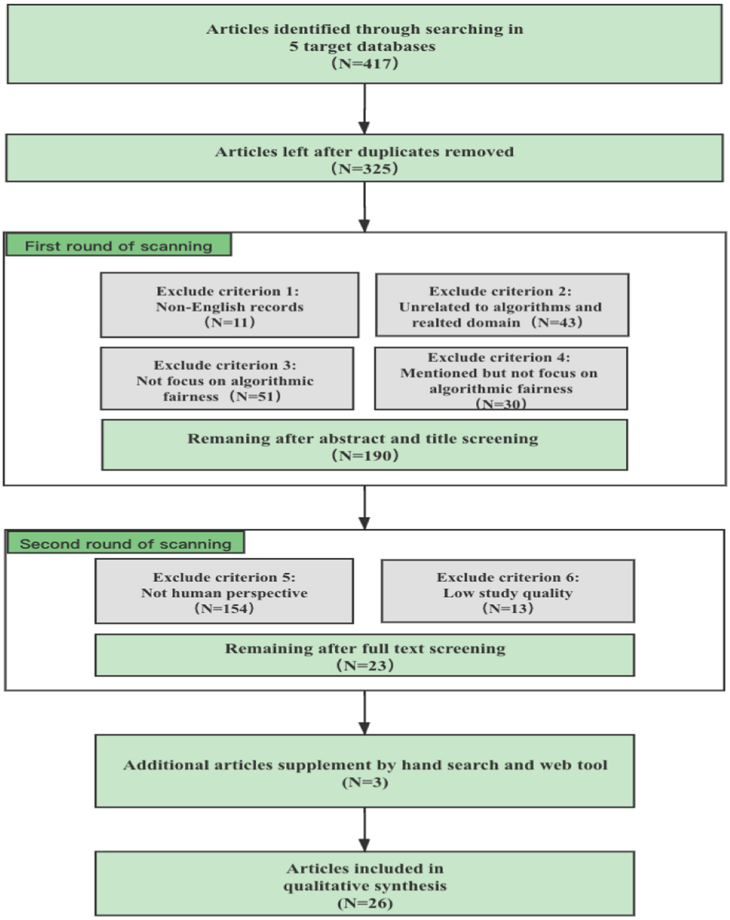

Fig. 1. Flow diagram of the studies selection.

3 Results

Through the search strategy developed earlier, we have retrieved 417 articles in 5 target databases distributed as follows: WOS 235, ACM digital library 93, APA 14, Wiley online library 33, and Elsevier ScienceDirect 42. Removing duplicates resulted in 325 articles.

After two rounds of screening based on the criteria we set, 23 documents were included. Indeed, 3 articles were included through web tools based on references provided in articles included before. Of these, 26 eligible articles were included. Figure 1 summarizes the process of study identification and selection.

Although 26 included articles seem to be relatively small, it's reasonable. Because the topic of algorithm fairness has only been widely concerned in recent years, the number of studies is generally tiny. What's more, studying algorithmic fairness from the human perspective is a branch of algorithmic fairness, which is a new research topic.

The reasons we use 26 articles for this systematic review are: 1) It has high value to conduct a systematic review of algorithmic fairness from a human perspective due to increasing attention to it. It's worth analyzing their content. 2) Research on this topic is still in its infancy. It is necessary to conduct a review to form systematic cognition to help with follow-up research.

Besides, [14] and [15] conducted their systematic reviews based on 25 and 19 included articles. This means conducted a systematic review with a small number of included articles is practical.

3.1 General Information About the Papers

We reported the qualitative analysis of the included studies with publishing time and paper type.

(1) Publishing time. Although our time limit in the search process is 2000–2020, the included articles' year of publication ranged between 2017–2020. The number of articles published from 2017–2020 are 3, 4, 13, 7. The time distribution of included article could indicate that HAF is a relatively new topic of algorithmic fairness. And there is also an increasing trend in the number of articles in 2017–2019, which means that more and more attention is being paid to algorithm fairness related topics from the human-centered perspective.

(2) Paper type. There are 20 conference papers, like ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) and The International World Wide Web Conference (WWW), there are 6 journal papers, and 4 of them were published in 2020. It reflects the transformation of outcome in HAF research from conference to journal, indicating that the relevant research gradually matured.

3.2 RQ1: Topics and Schemes

Through a content analysis of the 26 articles included in this review, we summarized the research topics of each article and finally excavated 12 research topics (some articles contain more than one topic, and there is an uneven distribution of these topics.). We sorted these topics into 4 categories to present the scheme of HAF research: phenomena and sources of algorithmic bias, users' perception of algorithmic fairness, promoting algorithmic fairness, the related concepts of algorithmic fairness (see Fig. 2). These topics indicated the research scheme of HAF research: social phenomenon appears → question perception → question solution → topic expansion.

- **Category I: Phenomena and sources of algorithmic bias**

The social demand for algorithmic fairness comes from the discovery and confirmation of algorithm discrimination/bias phenomenon. The exploration of algorithmic fairness begins with the study of algorithm discrimination/bias. Topics in this category include phenomena of algorithmic bias and sources of algorithmic bias.

Phenomena of algorithmic bias were detected by data collection and analysis in specific situations. Studies focus on this topic explore both intentional or unintentional bias in algorithm usage. Scholars revealed the phenomenon of algorithmic in multiple situations, like advertising [17], social media [21, 34], justice [26] and so on. This topic does not appear as the only topic of the article, but often with sources detecting algorithmic bias or promoting algorithmic fairness together.

Studies detecting sources of algorithmic bias usually analyze the causes of algorithm bias. Although different studies express the source of algorithm bias from different aspects, there are three main sources: 1) data collection bias. Engineers may use biased data during algorithm development without taking data representation into account [20]. 2) bias in data labeling. Some biases from the real world were brought into the data labeling process [36]. 3) attributes selection. Programmers may use attributes that might lead to bias, like population attributes, as they develop algorithms [24].

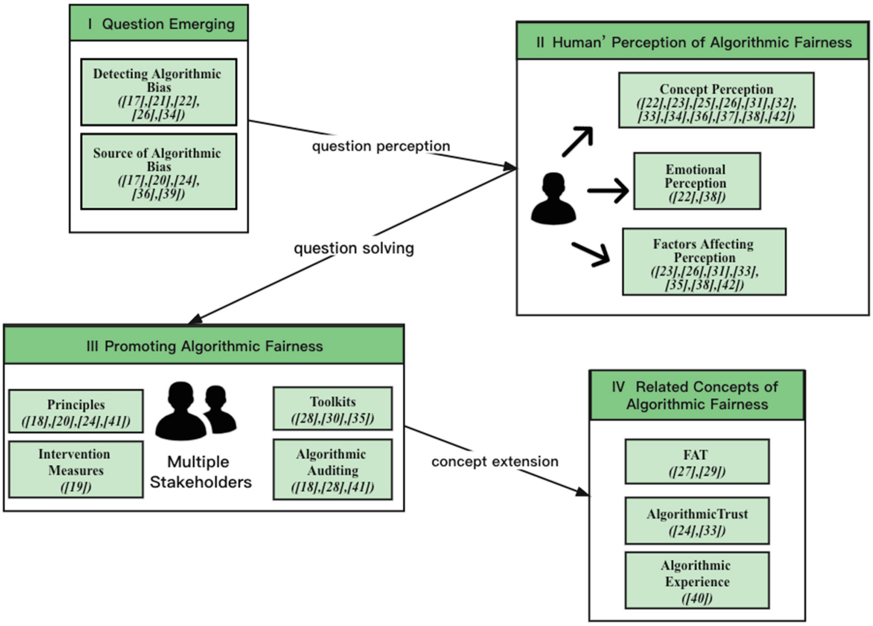
I Question Emerging
Detecting Algorithmic Bias ([17],[21],[22],[26],[34])
Source of Algorithmic Bias ([17],[20],[24],[36],[39])
question perception
II Human' Perception of Algorithmic Fairness
Concept Perception ([22],[23],[25],[26],[31],[32],[33],[34],[36],[37],[38],[42])
Emotional Perception ([22],[38])
Factors Affecting Perception ([23],[26],[31],[33],[35],[38],[42])
question solving
III Promoting Algorithmic Fairness
Principles ([18],[20],[24],[41])
Intervention Measures ([19])
Multiple Stakeholders
Toolkits ([28],[30],[35])
Algorithmic Auditing ([18],[28],[41])
concept extension
IV Related Concepts of Algorithmic Fairness
FAT ([27],[29])
AlgorithmicTrust ([24],[33])
Algorithmic Experience ([40])

Fig.2. Topics and scheme of HAF research.

- **Category II: Human' perception of algorithmic fairness**

The human's understanding of fairness may deviate from technical knowledge. Some scholars have studied algorithm fairness from human perception. There are three topics in this category: concept perception, emotional perception, and influencing factors of human perception.

On topics of concept perception, scholars summarize the popular definition of algorithmic fairness from the technical perspective. Participants are allowed to choose concepts they regard as fairer in experiments or questionnaires [22, 36]. Articles in this topic have a context characteristic, which means specific scenarios are usually setting for users to understand. For example, [25] set a credit loan scenario in their study for interviewees to make a loan decision.

Emotional perception refers to the emotional reflection of people's understanding of different algorithmic fairness definitions. For instance, [38] found that people with the strongest comprehension of demographic parity express the most negative sentiment toward it.

Influencing factors of human perception explore factors that influence human judgment of algorithmic fairness. Scholars have studied factors like the definition style of algorithmic fairness [23], personal experience of algorithmic bias [33], demographic characteristics [42], and so on. Especially, Wang et al. [35] explored situation factors and found that participants consider accuracy more important than equality when stakes are high.

- **Category III: Promoting algorithmic fairness**

This category mainly focuses on exploring measures to promote algorithmic fairness. Articles in this category were characterized by taking multiple stakeholders into account, including users, experts in related fields, industry practitioners, programmers, and so on. There are four topics in this category: developing principles for fair algorithm development, interventions of algorithm development, developing toolkits for algorithmic fairness, auditing of the algorithm.

The topic of developing principles for fair algorithm development aims to guide the practitioners to develop a fair algorithm by moral constraints. Some principles, like require engineers to select representative data and beware trade-offs, were put forward [20]; Interventions of algorithm development refer to giving developers some tips or help to assist engineers in developing fair algorithms. For example, [19] found that both representative data and interventions effectively promote algorithmic fairness based on control experiment; Some stakeholders suggest that toolkits for algorithmic fairness are needed, like processes and tools for fairness-focused debugging [28]; Besides, conducting algorithmic audits has also become recognized by stakeholders. They call for more proactive auditing processes and more holistic auditing methods [28].

- **Category IV: Related concepts of algorithmic fairness**

Articles in this category linked algorithmic fairness to concepts like algorithm trust, FAT (Fairness, Accountability, and Transparency), and algorithmic experience. Although topics of this category are scattered, algorithmic fairness is generally linked to other related concepts, which is beneficial for exploring algorithmic fairness.

Algorithm trust is a concept that refers to whether human trust algorithms, AI, or machine learning technics. As far as algorithm trust is concerned, studies were conducted to explore the relationship between algorithmic fairness and algorithm trust [27, 29]. [33] claim that algorithmic fairness might affect users' trust in algorithms and even affect their attitude towards companies and their products; FAT topic explores algorithmic accountability and transparency, bringing algorithmic fairness into a larger concept [40]. Conference on Fairness, Accountability, and Transparency have been held since 2018 to discuss FAT issues. Usually, scholars regard transparency and accountability of algorithms as an essential part of promoting algorithm fairness [24]; Algorithm experience refers to how users experience and perceive algorithms. Shin, [40] regarded algorithmic fairness as an essential index to measure users' algorithm experience.

3.3 RQ2: Methodologies Identified

We usually classify research methods into qualitative and quantitative methods. However, to answer RQ2, we noticed that acquiring human behavior and thoughts in human interaction with algorithms is one of the foundations of studies to explore algorithmic fairness from the human perspective. Based on the mindset of human-centered, we summarize the research methods of included articles from the two original dimensions: behavior collection and thought collection (see Table 1).

Table 1. Methods of HAF.

Dimension	Methods	Definition	Advantages
Behavior collection	Interventional behavior collection	Collecting behavior data in a laboratory environment	- Systematic observation of the participants' behavior - Control experiments have great flexibility
	Non-interventional behavior collection	Collecting behavior data through technical means without being detected by participants	- Observing participants in a natural state - Reduce the interference of external factors
Thoughts collection	Independent self-expression	Participants independently reported their personal feelings and thoughts	- Understanding user's authentic response by individual self-report - Participants think without influence from others

(*continued*)

Table 1. (*continued*)

Dimension	Methods	Definition	Advantages
	Group heuristic expression	Participants share their personal thoughts with other participants	- Participants could inspire each other about the topics - Opinions from different perspectives can be collected in a group

In the dimension of behavior collection, the methods used include interventional behavior collection and non-interventional behavior collection. Interventional behavior collection usually carries out laboratory experiments, setting scenarios (i.e., Criminal justice, credit lending, medical treatment, etc.) to guide the participants to complete relevant tasks and collecting the behavior data during them complete experimental tasks. Based on the characteristic of algorithmic fairness, scholars also add some technics in computer science into user experiments, such as setting up the user interface (UI) [20, 30], using algorithms to calculate the minimum number of tasks required for different participants [32], using some material made by algorithms [36] and so on. These technics could help participants record their choices and understanding of algorithmic fairness in a timely manner. The use of these new intervention methods provides new ideas for collecting user behavior in laboratory environments. Non-interventional behavior collection allows scholars to collect user behavior data in a natural state of participants. This method is often used to collect data on mainstream social media, such as Twitter and YouTube. Scholars use API to collect data to detect algorithmic discrimination [21, 22] and understand users' attitudes towards algorithmic fairness [37].

When it comes to thoughts collection, methods used in HAF research including independent self-expression and group heuristic expression. Independent self-expression collects individual self-feedback through questionnaires or one-on-one interviews. This method is generally used to express the participants' understanding of algorithmic fairness [23] and their opinions on promoting it [28]. Group heuristic expressions often use workshops or focus groups which allow participants to discuss as a group. This method could be used in studies which focus on a particular group, like potential affected by algorithmic bias [33]. It also is used to explore the measurements of promoting algorithmic fairness from the multi-stakeholder perspective [18, 40].

3.4 RQ3: Research Patterns of HAF Based on the Process of Algorithm Development

To answer RQ3, we try to conclude research patterns to explain what's human could do in the whole process of algorithm development. [16] concluded three types of mechanisms to enhance fairness in machine learning algorithms based on the timeline. Inspired by this, we summarized the general process of algorithm development first, including algorithm research, algorithm design, and algorithm application. And then we concluded 3 research patterns of HAF: pre-process pattern, in-process pattern, and post-process pattern (see Fig. 3).

Pre-process pattern focuses on the stage before algorithms are developed. It is necessary to fully investigate the algorithm on the applicable scenarios and target users before developing it. Pre-process pattern aims to get a comprehensive understanding of algorithmic fairness in a specific context. The object of these studies is the user of algorithm. It focuses on algorithmic bias detection, users' perception of algorithmic fairness definition, and their fair needs.

In-process pattern focuses on finding measures to promote algorithmic fairness during the process of algorithm design. Studies in this pattern usually focus on the staff involved in the development of the algorithm. For example, [19] tested the impact of interventions on the fairness of algorithms by adding interventional measures in the laboratory experiment.

Post-process pattern refers to the research pattern of testing the adaptability and fairness of users after designing an algorithm. This pattern aims to identify problems through actual usage.

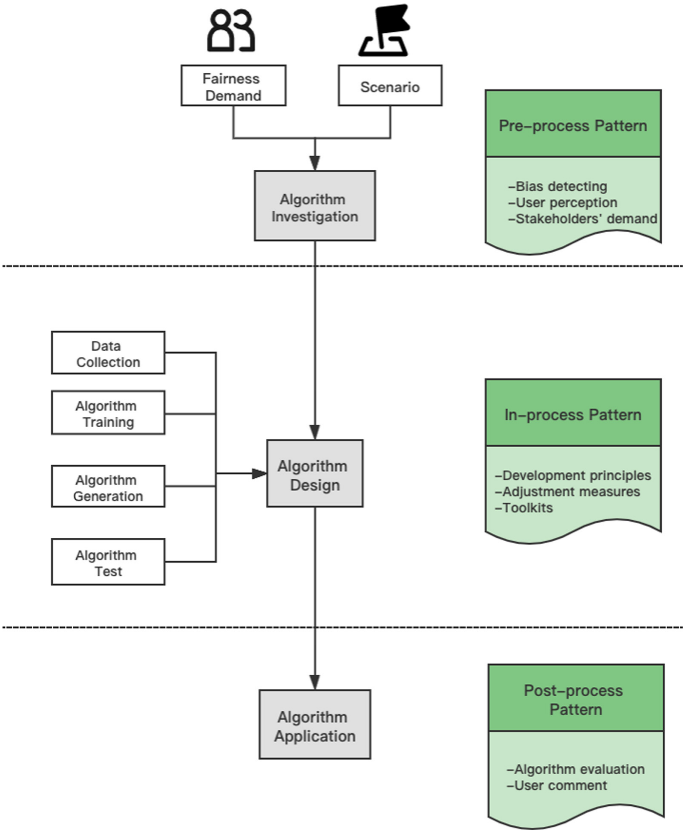

Fig. 3. Research patterns of HAF research.

4 Discussion

In this section, we focus on the relationship between these findings and future research, trying to find the research gaps and summarizing recommendations and implications for future research.

4.1 Taking Advantage of Human Research

Technical research and social research are complementary to each other. Technical research promotes algorithmic fairness through dataset testing, metrics conditioning, and model training. However, research from a human perspective has many advantages. One of the most common advantages is the ability to deeply understand the behavior and ideas of the algorithmic service recipients. Compared to deal with data, direct face-to-face contact with users through experiments or interviews could provide a more direct and in-depth understanding of users' needs and avoid the potential risk of algorithmic bias. Another advantage is that human research makes it easier to collect feedback in real-world use to assist programmers in finding problems. It will reduce the times of adjusting parameter for programmer. Therefore, it is necessary to integrate a human perspective in the whole process of algorithm development and conduct human-centered algorithm fairness research.

In Sect. 3 we summarized three pattern of HAF research. Future research could be conducted with this in mind, and here are some recommendations based on the three patterns.

Pre-process Pattern: Future research could explore human perception and demand for algorithmic fairness through more extensive questionnaires or interviews. Although the perception of algorithm fairness is variable in a different group of people, some universal demand could be extracted. Based on this, scholars could construct an evaluation index of algorithmic fairness from user and algorithm developer perspectives independently and achieving the standard expression and collection mode of the fair demand of the algorithm.

In-process Pattern: This pattern is mainly focused on the developers and practitioners in the field of algorithms. Some achievements of algorithmic fairness have been made at the technical level. These outcomes could combine with user research in the future, which could be more practical. For example, scholars could conduct user experiments to test different toolkits, adjustment measures designed by technical scholars and further summarized a systematic solution for promoting algorithmic fairness.

Post-process Pattern: Future research could collect large-scale user feedback for real-life usage scenarios, which depends on deeply explores the relationship between user behavior and user algorithm fair perception. Of course, it is also possible to capture the direct expression of the user through user experiments. The advantage of the experiment is that scholars could compare users' perception change between the pre-use stage and post-use stage, finding the gap between human need and actual usage and solutions for the gap.

4.2 Topic Expansion and Content In-Depth Mining

We analyzed 12 topics of HAF research and divided them into 4 categories which reflect the overall research scheme. It shows that HAF research has the characteristics of a wide range of topics and content coverage. However, we also noticed that HAF research has an uneven distribution of topics. The research gaps we found are as follows:

– Few scholars have systematically studied the consequences of algorithmic discrimination from the level of user behavior (e.g., burnout behavior, deprecation behavior, etc.).
– Current studies focus on the concept perception of a few special contexts like hiring, criminal justice, and medicine. In the future, more attention should be paid to context perception. The research scenario needs to be expanded. Therefore, we believe that Category II should focus more on contextual perception rather than concept perception in future studies.
– Current studies have explored users' perceptions of algorithmic fairness and its influencing factors. These factors are mainly focused on demographic characteristics. More factors and the relationship between factors still need to be discovered.
– The measures to promote algorithmic fairness are relatively fragmented, and no systematic solution has been developed.

Based on these gaps, we give some research recommendations from the topic level (see Table 2).

Table 2. Research recommendations of HAF research.

Category	Topic expansion	Detail
Category I: phenomena and sources of algorithmic bias	Human behavior when encounter algorithmic bias	The consequence of algorithmic bias: - Burnout behavior - Avoidance behavior - Deprecation behavior
Category II: human' perception of algorithmic fairness	Context perception	- Explore more practical context - Multi-contextual comparison
	Emotion perception	- Fine-grained emotional analysis - Connection of emotion perception and user behavior

(*continued*)

Table 2. (*continued*)

Category	Topic expansion	Detail
	Influencing factors of human perception	- Background of human (knowledge of algorithm, prior experience of algorithmic bias/discrimination) - Situational factors (risks of context) - Quantitative study of influencing factors - Interaction between the influencing factors
Category III: promoting algorithmic fairness	Theoretical construction	- Fusion of multi-disciplinary theory - The mechanism of system-pushing algorithm fairness
	HCI (Human Computer Interaction)	- UI design for algorithmic bias - User study of interaction between human and algorithm
	multi-stakeholders	- Gambling between multi-stakeholders - Finding systematic solutions
Category IV: related concepts of algorithmic fairness	Algorithm trust, FAT, Algorithmic experience	Deeper relationship exploration between each other

4.3 Physiological Signals: The Breakthrough Point of Research Methods

Research on HAF requires humans to be an important research object. This systematic review divided the research methods into four research methods from the two dimensions of human behavior collection and human thought collection in Sect. 3.

However, technological development has provided new support for the development of HAF research. In the future, scholars could also incorporate the collection of human physiological signals. Eye-tracking experiments and brain-computer interfaces can be used to collect more diverse data. For example, by collecting data like movement trajectory, gaze duration and frequency of eye-tracking, users' gaze interest areas could be found. It could help us to design user interfaces that cultivate users' algorithmic fairness perception. Besides, lectroencepha-lography (EEG) data could also be beneficial for scholars to explore user's cognitive burden of algorithmic fairness concept in different situations.

5 Conclusion and Limitation

As a socio-technical issue, algorithmic fairness not only requires technical exploration but also requires people to be included as important research objects. Algorithm fairness is a frontier topic in both computer science and social science, and HAF is a new trend in algorithmic fairness research. The articles included in this systematic review revealed that HAF research is gradually attracting more and more attention of scholars, and the number of articles is increasing year by year. Results show a trend from conference to periodical in article types. However, we also noticed that HAF research is still in the preliminary stage of exploration.

Our main contributions including (i) Presentation of an overview of existing research in exploring algorithmic fairness from the human perspective; (ii) Identified the research topics and scheme of HAF research, summarized the research methods based on the idea of human-centered, and concluded 3 research patterns; (iii) Concluded possible gaps in this field and summarized future research directions which could provide suggestions for future research.

There are, of course, limitations to our study. First, although we selected five mainstream databases, some articles that is written in English or not included in our target database may have been omitted. Second, limited by the topic of this review, we finally included only 26 articles for analysis, so this article could only provide a systematic overview of the initial stage of HAF research, which is time-sensitive.

Acknowledgement. This work is sponsored by Major Projects of the National Social Science Foundation: 19ZDA341.

References

1. Kleinberg, J., Mullainathan, S., Raghavan, M: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference. Berkeley, Article No. 43, p. 43:1–43:2 (2017)
2. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
3. Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. Proc. Priv. Enhanc. Technol. **2015**, 92–112 (2015)
4. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, pp.259–268 (2015)
5. Calders, T.: Verwer, S: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, Barcelona, pp.3315–3323 (2016)
7. Kallus, N., Mao, X., Zhou, A.: Assessing algorithmic fairness with unobserved protected class using data combination. In: Conference on Fairness, Accountability, and Transparency 2020, Barcelona, p.110 (2020)
8. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. Sociol. Methods Res. Article number: 0049124118782533 (2018)
9. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, pp.797–806 (2017)

10. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, pp.329–338 (2019)
11. Rosenbaum, H., Fichman, P.: Algorithmic accountability and digital justice: a critical assessment of technical and sociotechnical approaches. Proc. Assoc. Inf. Sci. Technol. **56**, 237–244 (2019)
12. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J. Clin. Epidemiol. **62**(10), 1006–1012 (2009)
13. Gough, D., Oliver, S., Thomas, J.: An Introduction to Systematic Reviews. Sage, Thousand Oaks (2016)
14. Tian, L., Kirsten, H.: Making professional development more social: a systematic review of librarians' professional development through social media. J. Acad. Librariansh. **46**(5) Article number: 102193 (2020)
15. Sørensen, K.M.: The values of public libraries: a systematic review of empirical studies of stakeholder perceptions. J. Doc. **76**(4), 909–927 (2020)
16. Pessach, D., Shmueli, E: Algorithmic Fairness (2020). arXiv:2001.09784 [cs.CY]

List of selected studies

17. Lambrecht, A., Tucker, C.: Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag. Sci. **65**(7), 2947–3448 (2019)
18. Koene, K., et al.: Algorithmic fairness in online information mediating systems. In: WebSci 2017, New York (2017). https://doi.org/10.1145/3091478.3098864
19. Cowgill, B., Dell-Acqua, F., Deng, S., Hsu, D., Verma, N., Chaintreau, A.: Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In: Proceedings of the 21st ACM Conference on Economics and Computation, Virtual Event, pp.679–681 (2020)
20. Rantavuo, H.: Designing for intelligence: user-centred design in the age of algorithms. In: Proceedings of the 5th International ACM In-Cooperation HCI and UX Conference, Indonesia, pp.182–187 (2019)
21. Salminen, J., Jung, S., Jansen, B.J.: Detecting demographic bias in automatically generated personas. In: Conference on Human Factors in Computing Systems 2019, Scotland (2019). https://doi.org/10.1145/3290607.3313034
22. Abul-Fottouh, D., Song, M.Y., Gruzd, A.: Examining algorithmic biases in YouTube's recommendations of vaccine videos. Int. J. Med. Inform. **148** Article number: 104385 (2019)
23. Dodge, J., Liao, Q.V., Zhang, Y.F., Bellamy, R.K.E., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp.275–285 (2019)
24. Veale, M., Kleek, M.V., Binns, R.: Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Canada, pp.1–14 (2018)
25. Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y.: How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, pp.99–106 (2019)
26. Grgić-Hlača, N., Redmiles, E.M., Gummadi, K.P., Weller, A.: Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: Proceedings of the 2018 World Wide Web Conference, Lyon, pp.903–912 (2018)

27. Williams, A., Sherman, I., Smarr, S., Posadas, B., Gilbert, J.E.: Human trust factors in image analysis. In: Boring, R. (ed.) AHFE 2018. Advances in Intelligent Systems and Computing, vol. 778, pp. 3–12. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94391-6_1
28. Holstein, K., Vaughan, J.W., Daumé III, H., Dudík, M., Wallach, H: Improving fairness in machine learning systems: what do industry practitioners need. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, pp.1–16 (2019)
29. Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C.H: In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**, 611–623 (2020)
30. Zhang, Y., Bellamy, R.K.E., Varshney, K.R.: Joint optimization of AI fairness and utility: a human-centered approach. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, pp. 400–406 (2020)
31. Loukina, A., Madnani, N., Zechner, K: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, pp.1–10 (2019). https://doi.org/10.18653/v1/W19-4401
32. Srivastava, M., Heidari, H., Krause, A: Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, pp.2459–2468 (2019)
33. Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warshaw, J: A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, pp.1–14 (2018)
34. Eslami, M: Understanding and designing around users' interaction with hidden algorithms in sociotechnical systems. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW), Portland, pp.57–60 (2017)
35. Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., Qu, H: Visual analysis of discrimination in machine learning. IEEE Trans. Vis. Comput. Graph. **27**(2), 1470–1480 (2020)
36. Barlas, P., Kyriakou, K., Kleanthous, S., Otterbacher, J.: What makes an image tagger fair. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, Larnaca, pp. 95–103 (2019)
37. Burrell, J., Kahn,Z., Jonas, A., Griffin, D: When users control the algorithms: values expressed in practices on the Twitter platform. In: Proceedings of the ACM on Human-Computer Interaction, Article number: 138 (2019). https://doi.org/10.1145/3359240
38. Saha, D., Schumann, C., McElfresh, D.C., Dickerson, J.P., Mazurek, M.L., Tschantz ICSI, M.C: Human comprehension of fairness in machine learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York (2020). https://doi.org/10.1145/3375627.3375819
39. Johnson, G.M: Algorithmic bias: on the implicit biases of social technology. Synthese (2020). https://doi.org/10.1007/s11229-020-02696-y
40. Shin, D., Zhong, B., Biocca, F.A: Beyond user experience: what constitutes algorithmic experiences. Int. J. Inf. Manag. **52**, Article number: 102061 (2019)
41. Lee, M.K., Kim, J.T., Lizarondo, L: A human-centered approach to algorithmic services: considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, pp. 3365–3376 (2017)
42. Pierson, E.: Demographics and discussion influence views on algorithmic fairness (2018). arXiv:1712.09124 [cs.CY]