Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects

Benjamin Laufer
Cornell Tech
United States
bdl56@cornell.edu

Sameer Jain*
Carnegie Mellon University
Pittsburgh, PA, United States
sameerj@andrew.cmu.edu

A. Feder Cooper*
Cornell University
Ithaca, NY, United States
afc78@cornell.edu

Jon Kleinberg
Cornell University
Ithaca, NY, United States
kleinberg@cornell.edu

Hoda Heidari
Carnegie Mellon University
Pittsburgh, PA, United States
hheidari@cmu.edu

ABSTRACT

Fairness, Accountability, and Transparency (FAccT) for socio-technical systems has been a thriving area of research in recent years. An ACM conference bearing the same name has been the central venue for scholars in this area to come together, provide peer feedback to one another, and publish their work. This *reflexive* study aims to shed light on FAccT's activities to date and identify major gaps and opportunities for translating contributions into broader positive impact. To this end, we utilize a *mixed-methods* research design. On the qualitative front, we develop a protocol for reviewing and coding prior FAccT papers, tracing their distribution of topics, methods, datasets, and disciplinary roots. We also design and administer a questionnaire to reflect the voices of FAccT community members and affiliates on a wide range of topics. On the quantitative front, we use the full text and citation network associated with prior FAccT publications to provide further evidence about topics and values represented in FAccT. We organize the findings from our analysis into four main dimensions: the *themes* present in FAccT scholarship, the *values* that underpin the work, the *impact* of the contributions both within academic circles and beyond, and the practices and informal norms of the *community* that has formed around FAccT. Finally, our work identifies several suggestions on directions for change, as voiced by community members.

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

FAccT, mixed methods, reflexivity, community perspectives, topics, values, impact

ACM Reference Format:
Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3531146.3533107

*Equal contribution, ordered by seniority

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
2022. ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533107

1 INTRODUCTION

"Fairness, Accountability, and Transparency" is a burgeoning area of research that examines the values embedded in socio-technical systems [3]. The area emerged amid concerns about the growing use of Artificial Intelligence (AI) and Machine Learning (ML) in socially consequential domains, and has evolved to include conferences, workshops, and books dedicated to this triptych of values. With this growth has come the responsibility of the research community to contribute positively to broader social change [2]. The present work offers a reflexive view toward this scholarship, and attempts to provide a clearer picture of its emergence as a significant interdisciplinary field. In particular, we focus on the ACM FAccT conference, the flagship venue at the center of these research efforts.[1] We analyze FAccT's contributions and shortcomings, with an eye toward identifying fruitful directions for near-term improvements.

Some scholars have argued that the community of researchers affiliated with FAccT has already made significant contributions, particularly by fostering numerous interdisciplinary interactions and raising awareness of the social, moral, and legal implications of technological work (see, e.g., [6, 18, 48]). At the same time, there have been calls from within the FAccT community warning against several troubling trends—including a disproportionate focus on a handful of narrow topics (e.g., mathematical formulations of outcome fairness) at the expense of pressing challenges, such as AI governance [1, 7, 19, 21, 41, 44, 61]. We aim to form a more nuanced and contextualized understanding of these views by consulting a variety of sources related to both data and methodology.

Reflexivity in scholarly field formation. This study takes a *reflexive* stance towards scholarly inquiry [14]. Bourdieu [12, p10] describes reflexive scholarship as *"objectifying the subject of objectification,"* by which he means *"deploying all the available instruments of objectification"*—quantitative and qualitative empirical methods—to identify presuppositions, underlying values and assumptions. Building on this notion, the present work puts forward a detailed analysis of FAccT research trends and themes to date with the goal of providing a foundation for broader community-wide discussions on the role and direction of the conference and scholarship. Our work takes a step toward igniting such constructive conversations

[1] FAccT has played a leading role in shaping the research agenda and it offers a concrete and meaningful way of specifying the boundaries of our inquiry. See [54] for a discussion of this type of *boundary specification problem* in other social-science research domains.

by reflecting the collective voices of FAccT community members—including their views on the past and their recommendations for the future. The authors of the present paper are themselves affiliated with FAccT, and in subsequent sections also reflect on the roles their own backgrounds play in the work. As such, this analysis may serve as an example of reflexivity for young fields of scholarship [13]. This work is motivated in part by calls for data reflexivity from inside the FAccT community [60] as well as in the context of ML [25]. Our work examines four critical pillars of FAccT: research **themes** embodied in its publications, **values** underlying the scholarship, **impact** (both intellectual and societal) that the work has garnered to date, and the practices and informal norms of the **community** of scholars who have come together through the venue. We ask:

- **Themes (Section 3):** What are FAccT's main research topics and subtopics? Which research approaches, methods and datasets are frequently utilized to examine these topics? Which topics have been studied more or less frequently? Does the community have sufficient access to 'high quality' datasets?
- **Values (Section 4):** What are the values underpinning FAccT publications? How has the community interpreted fairness, accountability, and transparency? Are there social values, moral foundations, ethical principles, and political ideologies that FAccT can address more extensively in the future?
- **Impact (Section 5):** To what extent have FAccT publications generated real-world and scholarly impacts? Has FAccT's intellectual influence been local (e.g., within intellectual echo chambers) or broad (e.g., bridging disciplinary gaps)?
- **Community (Section 6):** How do affiliates assess FAccT as a scholarly community? How do they perceive the practices, informal norms, and academic culture of the FAccT community?

We utilize a mix of qualitative and quantitative methods to elicit answers to the above questions (Section 2). From these analyses we obtain several insights: Our **thematic** exploration of FAccT publications shows that there has been an out-sized focus on ML as a computing subfield, and in quantitative work on fairness, displacing discussions about broader AI policy and governance (Section 3). We identify a list of **values** beyond fairness, accountability and transparency that are currently less well-represented at FAccT (Section 4). While our analysis does not attempt to describe the set of end-to-end deployments in practice, we report community views on deployment and broader **impact**, which are generally believed to be less than expected. We do find, however, that FAccT has played a positive role in exposing its community to insights and ideas from other disciplines (Section 5). Lastly, the **community**'s self assessment additionally reveals concerns around the practices and informal norms of the conference, including peer review practices, lack of inclusivity, and out-sized industry connections (Section 6).

Recommendations. To overcome some of the identified challenges, FAccT affiliates suggested several steps, including: (1) The conference organization should (perhaps ironically) be more transparent, especially concerning relationships with industry and aspects of the peer review process. (2) The community needs to foster a more inclusive environment, including (but not limited to) appreciating various modes of inquiry and forms of contributions. (3) The researchers must work closely with stakeholders and practitioners to have positive, real-world impact. (4) The scholarship must critically evaluate the assumptions it takes for granted, for example, by developing a more standard set of terms and norms that do the necessary translational work for the conference and make its communications more effective.

Table 1: Overview of our mixed-methods and datasets

Approach	Method	Dataset	Paper section(s)
Qualitative	Manual coding	FAccT pubs. corpus, survey responses	Themes (3), values (4), community (6)
Qualitative	Survey	Responses to open-ended questions	Values (4), impact (5), community (6)
Quantitative	Survey	Responses to multi-choice questions	Impact (5)
Quantitative	Network analysis	Citation network of FaccT articles	Themes (3)
Quantitative	Topic modeling	FAcct pubs. corpus	Themes (3)

2 OUR MIXED-METHODS DESIGN

This study follows a mixed-method design, summarized in Table 1: We collect both qualitative and quantitative data related to our research questions, and we utilize qualitative and quantitative methods to analyze the resulting data. Our motivation for using mixed methods is two-fold: 1) we want to present a comprehensive picture of FAccT; drawing on both qualitative and quantitative methods helps achieve this, as each helps address limitations of the other. 2) our hope is for this study to offer actionable guidelines to FAccT; given the multidisciplinary background of FAccT scholars, it is appropriate to use methods that not only capture, but also reflect, this diversity. We begin by providing additional context regarding our use of qualitative methods and reflect on our roles as researchers. We present overviews of our methods (Table 1) in Section 2.1 and a detailed description of our survey design in Section 2.2.

Reflecting on our roles as researchers. In qualitative research, the researcher is the key instrument for gathering and making sense of data. So their background and motives (cultural, disciplinary, personal, ethical, strategic, or otherwise) play an essential role in shaping the direction and outcome of their research. As such, it is paramount that they reflect explicitly on the potential influence of their background, biases, and values on the research process [22]. Following this tradition, we next provide information about our backgrounds and reflect on our roles as researchers.

Our team consists of five researchers, two in senior and three in junior roles. Collectively, our team represents a variety of gender identities, ethnicities, and cultural, socio-economic, and national backgrounds. However, our team does not represent a broad range of political views. We all identify with liberal, progressive and/or left-wing (as opposed to conservative) ideals and values. We are currently affiliated with academic institutions in the U.S., and some

of us have worked in the industry in the past. Three of us have had sustained interest and involvement in the FAccT scholarship. We generally hold favorable views toward the conference and the affiliated research community. However, our past experiences and perceptions regarding the lack of diversity in topics, backgrounds, and politics represented at FAccT motivated us to undertake the current study. We believe our close affiliation with FAccT elevates our understanding of the research landscape, and our sensitivity and care toward the challenges faced by FAccT affiliates and stakeholders. The background and experiences described above have undoubtedly shaped our choice of research questions and our interpretations of the data. For example, as mentioned earlier, we initiated the study with the conviction that the community would benefit from a more diverse representation of issues, politics, and research paradigms. Our position and background have also provided us with various forms of access—to citation data from AMiner and Semantic Scholar, to conference proceedings, and above all, to FAccT community members who agreed to participate in our questionnaire and share their reflections and recommendations.

2.1 Overview of Methods

Coding is a method of organizing qualitative data "comprised of processes that enable collected data to be assembled, categorized, and thematically sorted, providing an organized platform for the construction of meaning" [75, p45]. A code is often a "short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of [...] data" [66, p3]. We use manual coding in our thematic analysis of FAccT publications, and in interpreting the responses of FAccT affiliates to our open-ended questions about shortcomings and recommendations. Further details about our coding protocols will be presented in the appropriate sections.

Topic modeling, particularly the Bayesian unsupervised learning technique of Latent Dirichlet Allocation (LDA) [9], is a popular and well-documented tool in natural-language processing for eliciting thematic information from a text corpus. LDA operates on a bag-of-words representation of text documents and enables us to model each FAccT paper as a distribution over topics, where each topic is a distribution over vocabulary words. A topic can be understood as a set of frequently-present, co-located words, based on which we can assign a semantically meaningful overarching topic label [17]. From our model, we determine which papers belong to which topics; this enables us to develop a quantitative understanding of prominent research themes in FAccT—year-over-year changes and across all four years.

Community detection in citation networks is an unsupervised quantitative technique that enables us to split a citation graph into subgraphs (called *communities*), such that the nodes within a subgraph have denser connectivity, share properties, or play similar roles within the graph. We use the Louvain community detection algorithm [10] to elicit communities within the extended FAccT citation network—comprised of FAccT publications and their immediate citation connections. By definition, the papers within a community exhibit a higher concentration of citation relationships than those across communities [30], so we expect them to roughly map to sub-areas of research. We analyze titles of the papers that appear in each community to assign an overarching thematic label to it. We contrast the outcome of the above approaches to the thematic exploration of FAccT in Section 3.

Surveys directly solicit data from the population of interest—in our case, FAccT affiliates. By asking both open-ended and multiple-choice, Likert-scale rating questions [56], our web questionnaire aims to gather candid, less-biased views from participants [63]. Open-ended questions, a mode of structured interview, invite qualitative research analysis stemming from *phenomenology*. This style of analysis tries to identify and clarify phenomena as they are experienced by individuals, rather than from an abstract or objective perspective [36]. Closed-ended questions, which can be categorized under *quantitative survey design*, are suitable to answer descriptive research questions about the relationship between variables of interest—in our case, FAccT's topics and impact. The next Section provides details about our survey design.

2.2 A Mixed-methods Survey of FAccT Community Members

An essential component of our analysis is a web-based survey designed to solicit FAccT affiliates' responses to questions in three broad categories: 1) views on FAccT scholarship and recommendations for future improvements; 2) intellectual merit and broader impact of several FAccT research topics; 3) broader impact of FAccT scholarship in several application domains. The main questionnaire was followed with an optional set of questions about the participants' background and affiliation with FAccT. The qualitative component of our survey contained four main open-ended questions:

- *"Are there any moral or social values (sufficiently distinct from Fairness, Accountability, and Transparency) that you believe FAccT scholarship should address in near future?"*
- *"What do you consider to be the most important criticisms of FAccT scholarship to date?"*
- *"How do you believe the FAccT conference can address the above issues and limitations in the near future?"*
- *"Please briefly describe how you believe "impact" should be defined for FAccT scholarship."*

Participants also had the option of sharing additional thoughts about FAccT, further information about their backgrounds and identities, and feedback about the questionnaire. Additional details about survey design can be found in Appendix D, and the full survey is included in Appendix H.

Ethical considerations. Our study was reviewed and approved by Carnegie Mellon University's Institutional Review Board (IRB) prior to its launch. Beyond the standard criteria required to pass institutional review (e.g., obtaining informed consent), we accounted for several additional considerations, including our own motives and our participants' goals and aspirations. First, our motivation for involving FAccT affiliates in our research was to reflect their voices and opinions about the role and future of FAccT. Second, throughout the process of designing and administering the questionnaire, we were keenly aware of our participants' care and investment in the FAccT community and scholarship. As discussed by Howard and Irani [43], this awareness heightened our sense of duty to ensure their meaningful, active, and sustained participation in the

research process. Toward these goals, we attempted to make the questionnaire more collaborative by asking participants about their preferred definitions of "impact" and "value"—two of the key concepts we hoped to evaluate through our study. Additionally, we provided several opportunities for free-form expression of general thoughts and opinions regarding FAccT. Participants also were given the option of continuing their interactions with us (e.g., by emailing the Principal Investigator of the study directly, filling out the feedback textbox on the questionnaire, or expressing interest in participating in one-on-one interviews). Third, we made the questionnaire anonymous by default to prevent biasing participant responses (e.g., in anticipation of their responses being read and interpreted by researchers and later shared with the broader community). Nonetheless, we provided the option of identifying themselves and/or their responses if a participant so wished, so that we could name and acknowledge their contributions to our research. Fourth, to be mindful of our participants' time, we made all questions optional, but we mentioned that we appreciated their input on as many of them as they believed they were qualified to answer. Finally, we weighed the possibility of compensating our participants; considering the nature of their contributions, we concluded that monetary compensation could be perceived as disdainful commodification on our part and bias the sample.

Population, sampling, and respondents' demographics. We compiled a list of candidate participants by combining publicly available data of FAccT main-track authors, reviewers, and organizing committee members for 2018–2021.[2] Out of the 918 individuals emailed, 60 self-selecting FAccT affiliates responded (6.5% response rate). Questionnaire respondents were asked questions about their affiliation with FAccT and demographic information. Among all participants, 44 (75%) identified with STEM expertise and 23 (39%) identified with humanities, social sciences and arts (HSA) expertise (some marked both). At least 36 have participated in the FAccT conference as attendees, 39 as authors, 12 as organizing committee members, and 36 as reviewers. 54 respondents characterized their political views, of which 67% marked Liberal, 0 marked conservative or libertarian, and all remaining political views were individually submitted, of which the most-common was 'socialist.' 23 (41.8%) of respondents said that they belong to a marginalized/disadvantaged group, and 32 (58.2%) stated they did not.

[2]This list did not represent all conference attendees, authors of rejected papers, and other contributors who were not in the conference proceedings.

3 FACCT RESEARCH THEMES AND TOPICS

This section describes our thematic investigation of FAccT scholarship. We utilized one qualitative method (manual coding) as well as two quantitative methods (topic modeling and citation network analysis) to extract themes and patterns in FAccT publications.

Data collection. Our data consisted of text documents–i.e., articles that have been peer-reviewed and published by the FAccT conference in 2018–2021. Data collection was straightforward: We downloaded the full conference proceedings through the ACM website on September 25, 2021, with free access through Carnegie Mellon and Cornell University's credentials. In all, we downloaded 224 papers, 186 of which were full-length proceedings articles (the rest are non-archival extended abstracts).

3.1 Coding to Identify Topics, Methods, and Applications

Our qualitative coding aimed at organizing papers to respond to the following questions: What are FAccT's main research **topics and subtopics**? How have research efforts been distributed among

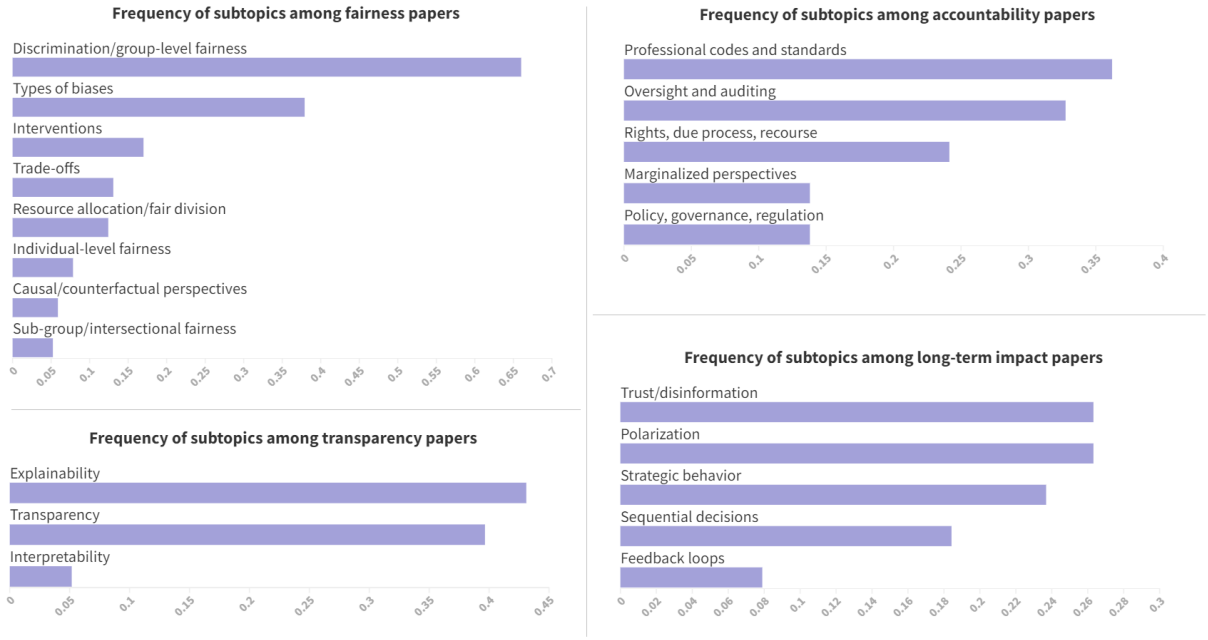

Figure 1: Relative frequency of subtopics among FAccT papers within 4 topics: fairness, accountability, transparency, and long-term impact.

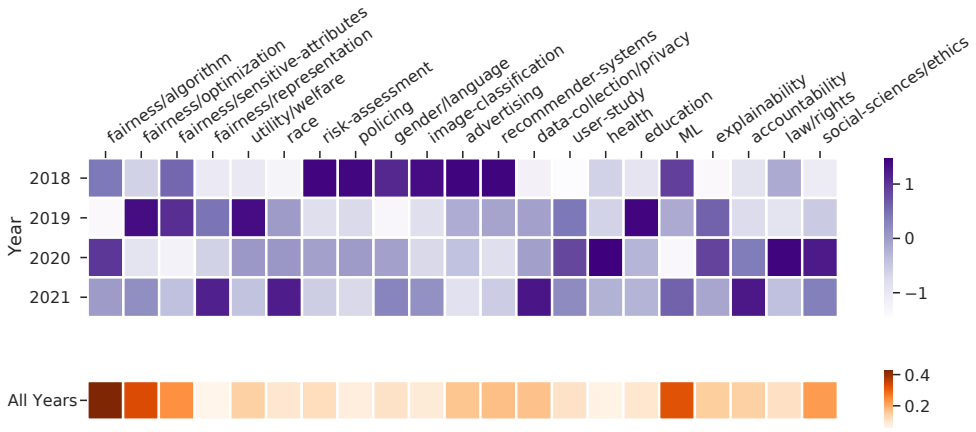

Figure 2: Topic distributions: normalized, grouped by year (top); across all years (bottom). By year, we see variations over time; e.g., `accountability` becomes more prevalent. For all years, we see the relative prevalence of topics in general. `fairness`-related topics and `ML` dominate all other topics.

them? Which research **approaches and methods** are frequently utilized? Which **application domains and datasets** have been studied extensively?

Coding process. We used a combination of *predetermined* and *emergent* codes [70]. Based on our research questions, our predetermined codes were classified into "research design", "topic", "application", and "data set". Within each category, we predetermined several codes as follows: Referring to standard classifications of quantitative and qualitative research approaches [22], we added 8 qualitative and 6 quantitative research designs as codes under the category of "research design" (see Appendix F). For "topics", based on our initial review of the data, we started with three high-level topics: `fairness`, `accountability`, and `transparency`. We used prior FAccT CFPs to determine subtopics under each of these broad topics. We went through 20% of papers in our dataset to validate our initial protocol. Accordingly, we added a fourth topic, `long-term impact`, as an emergent code. Our complete code book and additional details about our coding process are provided in Appendix F.

Findings. Our coding analysis revealed that in terms of the four broad research *topics* identified above, `fairness` has received the highest level of attention (69% of all publications), followed by `transparency` (26%) and `accountability` (26%), and `long-term impact` (17%). Other topics (e.g., `privacy` and `human factors`) were addressed in 31% of papers. Each paper can be categorized under multiple codes, so the percentages do not necessarily add up to 100. See Figure 1 for a sub-topic break-down of FAccT publications. `Discrimination/group-level`, `explainability`, `professional codes and standards`, and `trust/disinformation` were the most prevalent subtopics under fairness, transparency, accountability, and long-term impact, respectively.

In terms of *research design*, 18.9% of the papers used quantitative empirical methods (e.g., randomized experiments or causal methods) and 32.9% used qualitative empirical methods (e.g., interviews). 61.3% were broadly labeled as `STEM` papers and 29.7% were labeled as Humanities, Social sciences, and Arts (`HSA`). Further, we coded 11.3% of publications under `philosophy`, 11.3% as `professional`, and 9.9% as `law`. The top three *datasets* utilized in FAccT publications were `Adult Income` [50], `COMPAS` [53], and `German Credit` [42], all publicly available. Of the 75 papers identified as using an 'off-the-shelf' dataset, 18.7% used `Adult Income` 17.3% used `COMPAS`, and 10.7% used `German Credit`. A total of 23 papers (10.3%) used original, empirical datasets (excluding synthetic datasets).

3.2 Unsupervised Methods to Identify Themes

We employed two unsupervised approaches to discover themes across FAccT: 1) LDA-based topic modeling [9] on archival FAccT papers, and 2) community detection on a citation network consisting of FAccT publications and their immediate citation connections. In this section, we discuss the two approaches and draw insights from the two models. We provide additional details in the Appendix, and the accompanying code can be found in our online repository.

Topic modeling. We use Latent Dirichlet Allocation (LDA) [9] to elicit topics from FAccT proceedings. LDA enables us to model each paper as a distribution over (latent) topics, where each topic represents a distribution over vocabulary words. Higher-valued topic weights indicate that a topic is more prevalent in the learned model (Appendix A). We provide two heatmaps to visualize the learned topic distributions in Figure 2: topic weights grouped by FAccT conference year (top), in which we normalize each topic's weights to clarify changes year-over-year (i.e., we subtract a topic's mean over the 4 years and scale by the standard deviation), and topic weights for FAccT overall (bottom). We trained our model using $k = 22$ topics and initially examined the unnormalized results year-over-year. These results indicated that that one of the 22 topics contained words commonly used in the sciences to indicate uncertainty, which dominated over the other 21 topics; these were stop-word-like words like "may", "should", and "possibly." For clarity of presentation, we remove this topic from Figure 2.

Community detection. We use a variation of the Louvain community detection algorithm [10] to elicit *communities* within the citation network of all FAccT papers and their immediate neighbors, that is, papers that directly cite FAccT papers or are directly cited by a FAccT paper. (Recall that communities within a network often share common properties or play similar roles within the structure—in our case, they could help us identify research areas and topics.) We utilized two datasets for this analysis: the Semantic Scholar Open Research Corpus (S2ORC) [57] (Figure 3) and the AMiner citation network dataset [72]. Both sources provided incomplete citation data, so to ensure the robustness of our findings

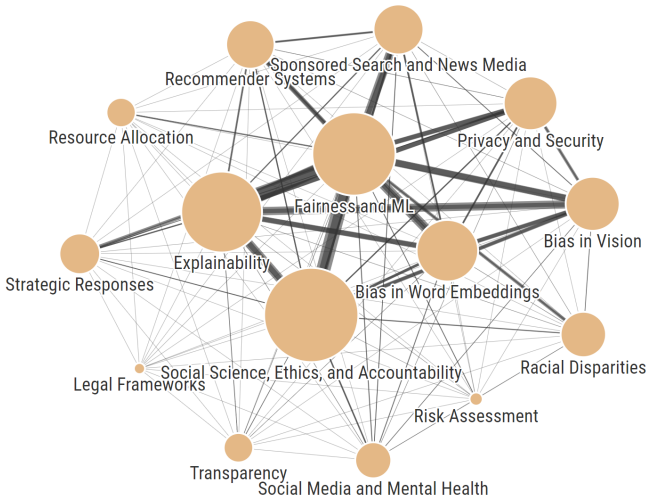

Figure 3: Communities detected using the S2ORC citation network. The size of a given node is representative of the number of papers in the corresponding community. Edge thicknesses represents the volume of citations between the respective communities.

Community Label	Top Terms and Papers
`Social Sc., Ethics, Accountability`	ethics social accountability Selbst et al. [68]
`Fairness and ML`	fair learning machine Hardt et al. [39]
`Explainability`	explanation learning model Ribeiro et al. [64]
`Bias in Word Embeddings`	bias embeddings word Bolukbasi et al. [11]
`Bias in Vision`	learning bias recognition Buolamwini and Gebru [16]
`Privacy and Security`	privacy data learning Dwork et al. [24]
`Recommender Systems`	recommender collaborative diversity Koren et al. [51]
`Sponsored Search and News Media`	search bias online Kay et al. [47]
`Racial Disparities`	race ethnicity disparity Hutchinson and Mitchell [44]
`Strategic Responses`	strategic signaling social Hardt et al. [38]

Table 2: The ten largest communities. Top terms are selected from the ten most frequent words in paper titles within each community. The referenced papers are the most-cited articles within each community.

to the idiosyncrasies of each data source, we analyzed both (Appendix B). The detected *communities* are visualized in Figure 3. Note that circles represent communitues, or subgraphs of the citation network with high concentrations of citation relationships. We named the communities by analyzing the titles of papers in each one (Appendix B).

Findings and discussion. Our topic model identified four fairness-related topics, which we label `fairness/algorithm`, `fairness-/optimization`, `fairness/sensitive-attributes` & `fairness-/representation`. Papers categorized under all four of these topics contain words like "fair" and "fairness," but are separable into distinct themes by the other words that comprise them. For example, `fairness/sensitive-attributes` contains attribute-class related words, including "parity", "protected", and "sensitive"; `fairness/optimization` has optimization-related words, such as "optimal" and "cost" (Appendix A). Notably, our results indicate that, separate from the well-documented fact that there are numerous, competing definitions of fairness [8, 19, 37, 41, 46, 49], there are also clear thematic differences in how FAccT papers examine fairness.

The overall results from our topic model (Figure 2, bottom) indicate that fairness-related topics and machine learning (ML) dominate the FAccT proceedings, both within and across years. This point is further supported by our citation analysis, where one of the largest communities we obtained corresponds to `Fairness and ML`, and other large communities correspond to bias in ML-related areas such as vision and natural language processing (NLP). The fact that we obtain these communities related to *both* fairness and ML (rather than separate communities pertaining to fairness and to ML) suggests not only that these topics are prevalent in FAccT papers independently, but also that fairness *for* ML is a dominant theme across FAccT.[3] We observe this result in spite of the more general way that FAccT defines its aims in terms of exploring "fairness, accountability, and transparency in socio-technical systems" [3], which importantly do not require the involvement of ML algorithms.

The focus on ML is also indicated by the `Explainability` community in the citation network. The `Explainability` community that we obtain focuses on ML models, and while it is less central to the citation network than the `Fairness and ML` community (as evidenced by the weaker strength of its connections with the other communities), both are similar in size. The focus on ML of both the fairness and the explainability communities is illustrated in Table 2, which shows that terms like "learning" and "models" are among the most frequently occurring terms in paper titles in those communities. Lastly, we also observe from the normalized topic distribution in Figure 2 (top) that there has been an increase in legal, social science, and ethics papers over the years; thus, while FAccT has predominantly concerned ML papers about fairness, there is evidence it has become increasingly expansive in scope over the years—so much so that, due to 2021, the overall-FAccT topic distribution in Figure 2 (bottom) marks this as one of the most prevalent themes. The citation network in Figure 3 confirms this result; however, it identifies a slightly different theme via a large catch-all community centered around `Social Science, Ethics, and Accountability`. In the same vein, the topic analysis also shows that accountability, though less of a focus than fairness, has increased in relevance each year from 2018-2021.

[3]For one illustrative example of how FAccT scholarship approaches this theme, see Hardt et al. [39], the top-cited paper in the Fairness and ML community.

4 VALUES UNDERPINNING FACCT SCHOLARSHIP

This section examines the values and principles underlying the existing FAccT scholarship. In particular, we ask how prior work has interpreted or operationalized fairness, transparency, and accountability as values. We also ask what other underlying values are represented in FAccT scholarship, and which ones potentially deserve further inquiry.

Interpretations of fairness, accountability and transparency. Within fairness, accountability and transparency, our coding of FAccT publications suggest that these concepts are multi-faceted and the scholarship has focused on specific aspects of these values (see Figure 1 for the breakdown of these topics into subtopics). In particular, the majority of articles addressing fairness have focused on manifestations of *outcome* disparities across socially salient *groups*. (This is, for instance, in contrast with the limited number of studies dedicated to *procedural* notions of fairness, or those that interpret it as proportionality or reciprocity). Existing work on accountability is mostly centered around *self-governance*—with the bulk of contributions proposing technical auditing tools or industry standards. Prior work has remained largely silent on *regulation* and mechanisms for enforcing compliance.

4.1 Values beyond Fairness, Accountability and Transparency

We asked our survey participants *"Are there any moral or social values (sufficiently distinct from Fairness, Accountability, and Transparency) that you believe FAccT scholarship should address in near future?"* 36 out of 60 participants responded.

Proposed values, meta-values, and related suggestions. Using manual coding to organize responses, we identified the following values (listed in alphabetical order) as welcome additions to FAccT's scope: Agency, benevolence, care, community service, dignity and respect, diversity and inclusion, freedom vs. oppression, harm prevention, intellectual property, loyalty, ownership, participation, privacy, reflexivity, reproducibility, safety, solidarity, sustainability, and truth vs. misinformation. Additionally, participants mentioned the need for broader conceptions of meta-values, such as equity, justice, and trustworthiness. For example, they brought up the necessity of:

- Scrutinizing structural facets of justice (e.g., the role of power and capitalism).
- Making room for non-Western values (e.g., indigenous values) and politics.
- Providing mechanisms for striking the right balance between conflicting values and interests (e.g. individuality vs. collectivity, corporate vs. government interests).
- Moving beyond principles and values toward practical implementation.
- Providing effective ethical education to AI experts.

Moral Foundations Theory (MFT). As an exploratory quantitative approach to understand the moral underpinnings of FAccT and identify commonly-held moral values currently not represented at FAccT, we report the results of an analysis in which we viewed FAccT publications through the lens of the *Moral Foundations Theory (MFT)*[4] [33]—a social psychological theory that describes the common bases of people's moral reasoning and tastes as considerations around several "foundations," including *care*, *fairness and reciprocity*, *loyalty*, *respect for authority*, and *sanctity and divinity*. Our analysis utilizes an existing computational tool associated with the MFT: the Moral Foundations Dictionary (MFD) [31, 34]. The MFD captures the core concepts and terms corresponding to each of the above five foundations. The frequencies of MFD terms appearing in the FAccT publications corpus along with the most commonly occurring terms under each foundation can be found in Table 3.

Table 3: The relative frequency of MFD terms in each foundation (out of 48,473 found terms) and the top-5 most frequent terms for each category in FAccT 2018-2021.

Foundation	Freq.	Most Frequent Terms
`care.virtue`	12.0%	health protect help share care
`care.vice`	4.7%	harm attack violent vulnerable victim
`fairness.virtue`	35.0%	fair trust equal law justice
`fairness.vice`	18.5%	bias discrimination disparity inequality unfair
`loyalty.virtue`	19.1%	group community company organization united
`loyalty.vice`	0.04%	outsider betrayed enemy heresy disloyal
`authority.virtue`	18.2%	order protect respect rank police
`authority.vice`	0.6%	illegal nonconformity unlawful refuse orders
`sanctity.virtue`	2.4%	body organic religion clean faith
`sanctity.vice`	1.4%	drug disease sexual mar pandemic

Findings and discussion. Table 3 provides suggestive evidence on the tendency of FAccT publications to be centered around the first two foundations of MFT, namely, *care* and *fairness*. Prior work has suggested that these two foundations are often emphasized in *liberal* cultures [34]. Note that while at a first glance foundations such as *loyalty* and *authority* may appear highly frequent in the FAccT corpus, a closer look at the words responsible for those numbers (i.e., 'group' and 'community' under *loyalty*, and 'order' and 'protected' under *authority*) suggests that their use in the FAccT context is likely in service of discussing fairness-related concerns. This observation should be contrasted with the concepts such as 'patriotism' and 'self-sacrifice' for the group or 'deference to legitimate authority' and 'respect for traditions' that underlie the association of these terms to *loyalty* and *authority*, respectively, in MFD. With that in mind, analyzing the terms within the MFD and MFT points to several additional values that are currently not a topic of inquiry at FAccT, including 'unity,' 'dignity,' 'spirituality,' 'divinity,' 'respect,' 'self-determination,' and 'freedom.'

[4]We emphasize that our usage of the theory is one among many possible approaches to explore the moral underpinnings of FAccT, and it should not be interpreted as us promoting it over other approaches to mapping the moral landscape (see, e.g., [28, 67, 69]).

5 THE INTELLECTUAL AND BROADER IMPACT OF FACCT SCHOLARSHIP

This section aims to characterize the *impact* of FAccT scholarship, in aggregate and across FAccT's topics and domains. To do this, we draw from the participants' responses to our survey. First, we code open-ended responses to our survey to characterize the FAccT community's notion of "impact" (Section 5.1). Second, we use Likert-scale responses to quantitative questions to understand views on the following impact-related topics: interdisciplinary exposure and peer review (Section 5.2), the impact and priority of various research topics (Section 5.3), and the impact of different application domains (Section 5.4). Finally, we discuss critiques put forward by respondents concerning impact (Sec. 5.5).

5.1 Defining Impact

This section reports the participants' views on the "intellectual merit" and/or "broader impact" of FAccT scholarship across a variety of topics and domains. For concreteness and grounding, we mentioned the standardized definitions of these terms proposed by the US government's National Science Foundation (NSF), which broadly defines *intellectual merit* as the contribution to advancement of knowledge and understanding, and *broader impacts* as benefits to society and contributions to the achievement of specific, desired societal outcomes. (Further criteria and examples can be found in Appendix C) We also provided participants with the opportunity to define impact. In particular, we asked them to *"briefly describe how [they] believe 'impact'* should *be defined for FAccT scholarship."* Many of our participants thought the NSF definitions are sufficiently broad. Others considered impact to be subjective and difficult to measure for the scholarship as a whole. The rest shared their suggestions on how impact should be defined. Utilizing manual coding, we categorized all responses. The resulting codes and relevant quotes can be found in Table 4.

5.2 Analysis of Interdisciplinarity

FAccT is a venue which features scholarship from a variety of disciplines and backgrounds. The cross-disciplinary topics and research questions that surface in FAccT scholarship require meaningful and constructive communication across these diverse perspectives. We leverage responses to two questions in our survey to characterize how successfully FAccT has facilitated and moderated such communications.

Findings. In response to our question, *"To what extent has the FAccT conference exposed its members to insights and ideas from other disciplines?"*, the mean Likert score across participants was 3.95 ($SE = 0.190, n = 58$), suggesting the general belief that FAccT has positively impacted interdisciplinary interactions.

However, one potential challenge for FAccT's interdisciplinary community is its peer review process. We asked participants, *"Reviewers of the FAccT conference come from various disciplines and backgrounds. In your experience, how has this impacted the quality of the conference's peer review?"*. Compared to the first question about exposure, respondents' assessments were less positive, with a mean score of 3.11 ($SE = 0.229, n = 55$). Notably, these answers did not seem to vary with participants' background or expertise.

Community suggestions. Participants offered several additional criticisms and suggestions concerning the interdisciplinary nature of FAccT. We classified their responses into three categories: 1) those pertaining to interdisciplinary communications 2) collaborations and 3) contributions. On the topic of communication, responses pointed out the need for **integrating non-STEM perspectives**. For example, one response pointed out a *"lack of integration between FAccT and philosophical community (including ethicists, political philosophers), due to the technical nature of FAccT scholarship."* On the topic of collaboration, one suggestion was to create an **award for collaborative scholarship**. Other participants expressed **concerns about quality and rigor** of contributions that

Table 4: Coding and categorization of FAccT affiliates' notions of impact.

Code/Category	Quotes
Enhancing public literacy and awareness	*"Improving awareness of [fairness, accountability, and transparency]-issues in the broader public; improving data & [AI] literacy of the broader public"*, *"[Shaping] the broader discourse around what gets built, why, and how."*
Educating AI experts	*"Ethical frameworks and education of those in the AI community"*
Policy influence	*"Real-world influence on [...] policies, [...] software development and data handling standards."*
Concrete benefits for impacted communities	*"Benefit to society from consideration of historically marginalized perspectives/experiences. Applicability of findings/frameworks/techniques to improving technology as deployed in the real world."*
Translation to practice	Example: Contributing *"open source scientific software"*
Buy-in from impacted communities	*"Enthusiastic participation and research support from communities and identities most likely to experience algorithmic harm under current practices."*
Intellectual paradigm shifts	*"[Changing] currently accepted conceptions, terminology, and frameworks"*, *"Reframing how key stakeholders (decision-makers, policy creators, advocates) understand/think about the world."*
Academic reach	*"Ability to reach different subcommunities in FAccT (CS, Econ, Sociology, etc)"*
Political ramifications	*"FAccT scholarship should not be misused by economic or political stakeholders to calm down upcoming discussions about AI ethics."*, *"A particularly negative impact is [the] uptake of technical approaches by industry in ways that amount to empty and detrimental solutionism."*

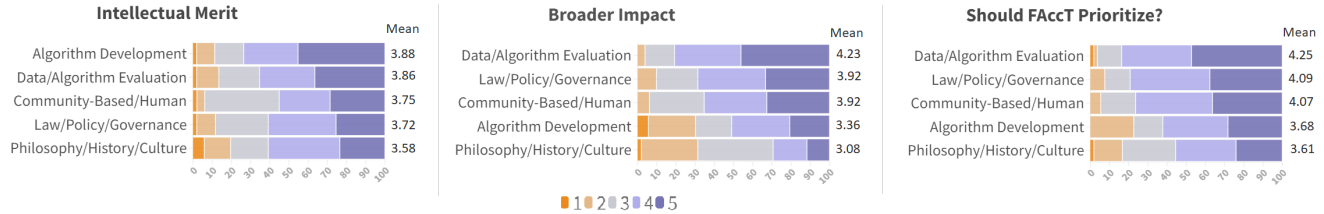

Figure 4: Participants' views on FAccT scholarship across five topics, specifically their intellectual merit (left), broader impact (middle), and the topics' priority for future FAccT scholarship (right). We illustrate the distribution of respondents' ratings on the 1-5 scale, with 1 corresponding to 'very low' and 5 corresponding to 'very high.' The percentage of participants who gave a particular rating is represented on the x-axis and mean ratings are reported on the right end of the graphs.

are considered interdisciplinary. One response stated, *"It is not uncommon for individuals to be exposed to a handful of papers, persons, and ideas that are not in their area of expertise, and then think they can produce new scholarship on that basis, because the field encourages 'interdisciplinarity.'"* Suggestions related to this criticism emphasize improving the peer review and other organizational aspects of the conference. For instance, one common suggestion was separating tracks and reviewing pools based on scholarly expertise. We discuss these ideas in further detail in Section 6.1.

5.3 Topic-level Analysis of Impact

Survey participants reported their assessment of intellectual merit, broader impact, and priority of FAccT scholarship pertaining to five main topics. The topics were drawn from FAccT's 'tracks' listed on the conference CFPs. Further details about how we chose these topics can be found in Appendix E. We draw a distinction here between *topics* and *application domains*, which we turn to in Section 5.4. In particular, topics are disciplinary, methods-based, or abstract, and they may pertain to any number of application domains. Topic-related responses are visualized in Figure 4.

Findings. We highlight several thought-provoking patterns in participants' responses. First, whereas algorithm development is perceived as a topic with relatively high intellectual merit, its broader impact and priority scores are significantly lower, lagging behind three other topics. Second, the philosophy/history/culture topic may appear as producing scholarship with comparatively lower merit, impact, and priority. However, we note that one potential explanation for this low score can be perceptions of better/more specialized venues for publishing philosophical contributions outside FAccT. (A similar sentiment came up around topics, such as privacy, during our think-aloud protocols). Related to the importance of philosophical contributions for FAccT, one participant pointed to problems of *ontology* affecting the quality and rigor of FAccT work.

5.4 Analysis of Domain-specific Impact

FAccT scholarship concerns several high-stakes application domains, ranging from criminal justice to education. Focusing on seven key domains which are prominent in FAccT, we asked respondents to characterize FAccT's broader impact. (We did not ask respondents to surmise the intellectual merit and priority of these domains, because most of them are known to be of high importance and have significant scholarly attention). Results are reported in Figure 5.

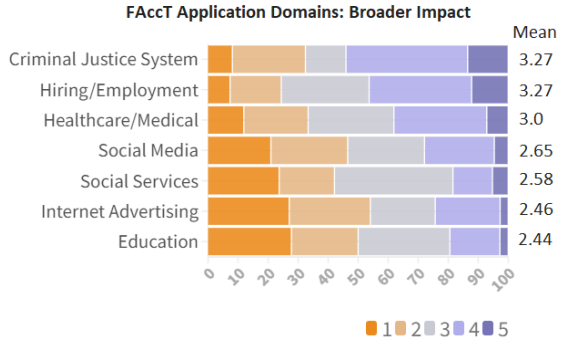

Figure 5: Application domains in order of the perceived broader impact of relevant FAccT scholarship (score of 1 corresponds to 'very low' and 5 to 'very high'). The percentage of respondents that gave a certain rating is represented on the x-axis and mean ratings are reported on the right. Criminal justice system and hiring/employment received the highest scores on average, whereas education and internet advertising received the lowest scores.

Findings. We see significant imbalances between respondents' ratings of one cluster of domains (criminal justice system, hiring/employment, and healthcare/medial) and the rest (social media, social services, internet advertising, and education). We also find that, even though education was deemed important enough to warrant its own CFP track in 2020, respondents found it to have relatively lower broader impact in comparison to the other domains.

5.5 Criticisms and Suggestions

The responses to survey questions reporting in Figures 4 and 5 suggest that there is a rift between the convictions held by community members and the broader impact of some of the work. To explain this rift and find strategies to overcome it, we draw from participants' open-ended answers about major criticisms and suggestions for FAccT. In particular, we identify a number of barriers to broader impact: (1) Insularity, (2) Narrow Inquiry, and (3) Ontology.

Insularity. A number of respondents pointed out a key shortcoming of FAccT as a failure to *"address actual problems with impacted communities."* One respondent connected this to **echo-chamber dynamics** within FAccT. Multiple respondents pointed out that FAccT papers overly adhered to **solutionism**, with one stating that

FAccT scholarship has *"a tendency to not fundamentally question certain technologies or discuss in what contexts they're (in)appropriate [...] instead [it exhibits] a tendency to incrementally improve such technologies."* To overcome these issues, respondents suggested **engagement with domain expertise** and **public and community engagement**. One respondent recommended that FAccT should *"emphasize partnering with translational researchers in real disciplines with 'street-level bureaucrats' and practitioners."* Another broad category of solutions concerned **diversity and inclusion**. This included a call to expand *"Diversity in authors, reviewers, and general participants"*, by direct outreach and financial assistance.

Narrow inquiry. A number of respondents pointed out that FAccT scholarship has had limited impact because its focus is too narrow or granular. One respondent pointed out the *"technical/technocratic approaches to fairness [...] offer a very narrow view of more broadly understood conceptions of fairness in law, policy, society."* Further, respondents warn against the phenomenon of **fair-washing**, where narrow notions of fairness get used to condone existing practices. Other responses pointed out an over-emphasis on **Machine Learning**, with one stating *"I don't think I've ever read a [FAccT] paper that didn't in some way have to do with ML."* Other responses pointed out the over-representation of **Western- and U.S.-centric** values in FAccT, which restricts the scope of applicability and impact of FAccT contributions. One respondent stated: *"The conversation in FAccT often assumes a strong familiarity with a set of values and modes of discourse which are currently highly visible in the US, but much less so in other countries, especially non-Western countries. I urge the community to consider what can be done to make researchers who are not fluent in this specific type of discourse feel welcome and able to participate in the community."*

Ontology. Another critical factor that may hinder FAccT's impact, according to participants, has to do with the conference's definition of **key concepts**. One respondent stated, *"FAccT has not yet settled key ontological questions about the field, leading to incoherent use of key terms, such as 'AI' and 'algorithm.'"* Another respondent noted a lack of **critical evaluations of mathematical assumptions**, stating that FAccT suffers from *"an over-focus on technical results without careful considerations for the mathematical assumptions and constraints used"*. One emergent suggestion encouraged *conceptual* work, in particular, drawing on Science and Technology Studies.

6 FACCT COMMUNITY REFLECTIONS

This Section aims to shed light on the FAccT *community*'s practices, inter-personal norms, and (academic) culture. The analysis provided here is based on the responses participants gave to our open-ended question inquiring about the most important *criticisms* of FAccT to date and *suggestions* for improvements. Responses covered a wide spectrum of opinions, ranging from broadly positive to highly negative, but the majority of responses lied in the middle of this spectrum–bringing up significant challenges and tensions while recommending steps for improvement.

Coding process. 42 (out of 60) participants responded to the relevant open-ended questions. We followed the process of manual coding proposed by Tesch [73] to extract patterns from the resulting qualitative data. The output of our coding process for all open-ended responses is summarized in Table 8 in the Appendix. At a high-level, our analysis revealed two major categories of worries concerning the FAccT *community*: one regarding the *organization* and role of FAccT as a conference (including its relationship with industry, government, and traditional disciplines) and another regarding the affiliates' attitudes toward each another and the resulting *culture*.

6.1 FAccT as a Venue

Participants scrutinized several organizational facets of FAccT as a conference. Major themes reflected in their responses included **peer review quality**, which was identified as particularly challenging for FAccT due to its interdisciplinary mission and the varied disciplinary expertise of FAccT reviewers. Another concern related to the publication process at FAccT was the risk of **over-curation** of accepted papers and tutorials, as opposed to *"accept[ing] all work that is novel, correct and fits the scope of the conference"*. Participants offered several suggestions to improve peer review quality. One proposal that appeared several times was **separating tracks and reviewing pools by expertise.** One respondent cautioned against this idea, though: *"FAccT should strive to be a space that transcends specific disciplinary standards and traditions rather than perpetuating them. We should make more effort to have reviewers from more disciplines rather than creating silos where papers are reviewed only by members of their own disciplines."* Another proposal was to foster a **more transparent reviewing process** by clarifying standards and quality measures, possibly revealing the expertise of the reviewers assigned to each submission, and devising an educational/on-boarding process for reviewers.

The second major theme regarded the **influence of industry** on FAccT as potentially troubling. For example, one participant warned that *"The FAccT community should be more careful in how researchers from big tech industry are assigned key positions in the conference."* Another described the potential harms of unregulated corporate influence as follows: *"It is easy for industry to adopt (or co-opt) some aspects of the work in ways which only minimally help those affected by algorithmic systems, and may even make them worse by giving them a false sense of legitimacy."* Corporate interests can additionally bias the scholarly discourse: *"Some questions are not asked or are difficult to ask within industrial research. There is a political bias."* As a way of moderating the influence, one participant called on FAccT to *"implement stricter funding disclosures for submitted manuscripts, limit corporate researchers' involvement with the [organizing committee], diversify the [organizing committee], promote CRAFT, [clarify that] FAccT is not an auditing organization."* Some respondents acknowledged the complex nature of FAccT's connections to industry, and hoped that it can *"figure out a way to enable industry to meaningfully contribute (there are many good researchers out there, despite the poor dynamics we've read about in the past year)."*

In addition to above concerns, respondents urged the community to reflect on the potential negative impact of FAccT on standard CS conferences (e.g., by *"isolat[ing] people who are concerned with values-oriented work from the main body of the technical community"*); the environmental harms of FAccT as a large academic gathering; and the risk or uncertainty of affiliation with FAccT for junior researchers.

6.2 FAccT's Interpersonal Norms and Practices

Responses concerning the interpersonal norms and culture of FAccT painted a picture of striking **divide between several major camps**: *STEM-focused*, *social-scientific*, and *activism-oriented* efforts. One participant described the tension as *"between those who see incremental progress as progress and those who believe the only meaningful progress will be revolutionary."* On one side of the spectrum, respondents criticized *"viewing research with a specific ideological prism rather than a scientific prism."* Some sensed outright **animosity toward STEM**: *"I found myself in several situations where people would speak with true disdain towards engineers, as though this was commonplace and normal."* The participant further expressed shock toward this attitude, especially *"for a conference about fairness, accountability and transparency."* The **lack of inclusivity** was reflected in calls for increased outreach and tolerance.

Perceptions of rampant **non-constructive critisism** were also common among participants. For example, one participant said *"FAccT scholarship is all about [criticism] now ... Do we want to be a community of people that* does *stuff or a community of people that* complains *about stuff?"* As a remedy, one participant proposed encouraging **constructive criticism** by dedicating to it a phase after peer review: *"once papers are [accepted], there should be another stage which is constructive criticism where people are invited to challenge the assumptions/values/agendas of the work"*.

7 CONCLUSION

This reflexive study aimed to shed light on the scholarly field forming around fairness, accountability, and transparency in socio-technical systems. We analyzed the FAccT conference—the *themes* present in its publications, the *values* that underpin the research, the *impact* of the work, and the culture of its *community*. Our mixed-methods analysis has used both quantitative and qualitative approaches to study how the FAccT community has directed its emphasis over the past several years. Our analysis highlights several significant needs and opportunities in the FAccT community, including (1) further intellectual investment in pressing issues of governance and accountability, (2) expanding the values underpinning the scholarship, (3) strengthening connections to real-world issues, practices, and stakeholders, and (4) building a more inclusive community. In closing, we hope our contribution benefits the FAccT community by facilitating a constructive dialog around the challenges we face as a diverse, interdisciplinary field aiming to address sensitive, high-stakes socio-technical issues that will only grow in magnitude and significance in the years to come.

ACKNOWLEDGMENTS

First and foremost, we acknowledge the vital role that survey participants played in our research. We sincerely appreciate their motivation to benefit the FAccT community by partaking in our study, and we are grateful for their time, mental energy, and commitment to the field. We are grateful to Karen Levy for multiple rounds of invaluable feedback at various stages of this project. The authors additionally would like to thank the participants of the AI, Policy, and Practice (AIPP) group at Cornell, and the Digital Life Initiative (DLI) at Cornell Tech for their thoughtful remarks and suggestions. Last but not least, we are indebted to Madiha Z. Choksi, Emanuel Moss, Emily Tseng, Meg Young, Nil-Jana Akpinar, Amanda Coston, Wesley Deng, Kit Rodolfa, and Joshua Williams for helping to workshop and develop the research instruments for parts of this project.

Benjamin Laufer is supported by DLI at Cornell Tech and NSF CNS-1704527. A. Feder Cooper is supported by AIPP at Cornell University, DLI at Cornell Tech, and the John D. and Catherine T. MacArthur Foundation. This material is based upon work supported in part by NSF IIS2040929, a Simons Investigator Award, and a grant from the John D. and Catherine T. MacArthur Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and other funding agencies.

REFERENCES

[1] J. Khadijah Abdurahman. 2019. A Response to Racial Categories of Machine Learning by Sebastian Benthall and Bruce Haynes. https://medium.com/@blacksirenradio/fat-be-wilin-deb56bf92539
[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Barcelona, Spain, 252–260.
[3] ACM. 2022. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT). https://facctconference.org/index.html.
[4] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The Effect of Differential Victim Crime Reporting on Predictive Policing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 838–849. https://doi.org/10.1145/3442188.3445877
[5] Maria Antoniak. 2021. little-mallet-wrapper. https://github.com/maria-antoniak/little-mallet-wrapper
[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org, website. http://www.fairmlbook.org.
[7] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 149–159.
[8] Reuben Binns. 2018. What Can Political Philosophy Teach Us about Algorithmic Fairness? *IEEE Secur. Priv.* 16, 3 (2018), 73–80. https://doi.org/10.1109/MSP.2018.2701147
[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. http://jmlr.org/papers/v3/blei03a.html
[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.
[12] Pierre Bourdieu. 2000. *Pascalian meditations*. Stanford University Press, Stanford, CA, USA.
[13] Pierre Bourdieu. 2004. *Science of science and reflexivity*. Polity, Cambridge, UK.
[14] Pierre Bourdieu and Loïc JD Wacquant. 1992. *An invitation to reflexive sociology*. University of Chicago press, Chicago, IL, USA.
[15] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. CRC press, Boca Raton, FL, USA.
[16] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
[17] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'09)*. Curran Associates Inc., Red Hook, NY, USA, 288–296.
[18] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* 1, 1 (2018), 1–13.

[19] A. Feder Cooper and Ellen Abrams. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 ACM Conference on AI, Ethics and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 46–54. https://doi.org/10.1145/3461702.3462519

[20] A. Feder Cooper, Maria Antoniak, Christopher De Sa, Marilyn Migiel, and David Mimno. 2021. 'Tecnologica cosa': Modeling Storyteller Personalities in Boccaccio's 'Decameron'. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, Punta Cana, Dominican Republic (online), 147–153. https://aclanthology.org/2021.latechclfl-1.17

[21] A. Feder Cooper, Karen Levy, and Christopher De Sa. 2021. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, New York, NY, USA, Article 4, 11 pages. https://doi.org/10.1145/3465416.3483289

[22] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE publications, Los Angeles, CA, USA.

[23] Nicolas Dugué and Anthony Perez. 2015. Directed Louvain: maximizing modularity in directed networks.

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.

[25] Madeleine Clare Elish and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication monographs* 85, 1 (2018), 57–80.

[26] K Anders Ericsson and Herbert A Simon. 1998. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity* 5, 3 (1998), 178–186.

[27] William P Eveland Jr and Sharon Dunwoody. 2000. Examining information processing on the World Wide Web using think aloud protocols. *Media Psychology* 2, 3 (2000), 219–244.

[28] Alan Page Fiske. 1991. *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press, New York, NY, USA.

[29] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. 1993. A description of think aloud method and protocol analysis. *Qualitative health research* 3, 4 (1993), 430–441.

[30] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486 (2010), 75–174.

[31] J. A. Frimer, R. Boghrati, J. Haidt, J. Graham, and M. Dehgani. 2019. Moral Foundations Dictionaries for Linguistic Analyses, 2.0. http://www.jeremyfrimer.com/uploads/2/1/2/7/21278832/summary.pdf.

[32] Stuart Geman and Donald Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6, 6 (1984), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

[33] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, New York, NY USA, 55–130.

[34] Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96, 5 (2009), 1029.

[35] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (2004), 5228–5235. https://doi.org/10.1073/pnas.0307752101

[36] Thomas Groenewald. 2004. A phenomenological research design illustrated. *International journal of qualitative methods* 3, 1 (2004), 42–55.

[37] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Barcelona, Spain, 501–512.

[38] Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. 2015. Strategic Classification. *CoRR* abs/1506.06980 (2015), 1–26. arXiv:1506.06980 http://arxiv.org/abs/1506.06980

[39] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., Barcelona, Spain, 1–9. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[40] Hoda Heidari and Jon Kleinberg. 2021. Allocating Opportunities in a Dynamic Model of Intergenerational Mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 15–25. https://doi.org/10.1145/3442188.3445867

[41] Deborah Hellman. 2020. Measuring Algorithmic Fairness. *Virginia Law Review* 106 (2020), 811.

[42] Hans Hofmann. 1994. German Credit Dataset.

[43] Dorothy Howard and Lilly Irani. 2019. Ways of knowing when research subjects care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Glasgow, United Kingdom, 1–16.

[44] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[45] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Mach. Learn.* 37, 2 (1999), 183–233.

[46] Maximilian Kasy and Rediet Abebe. 2021. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Virtual, 576–586.

[47] 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)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[48] Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, New York, NY, USA.

[49] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807* 1, 1 (2016), 1–23. http://arxiv.org/abs/1609.05807

[50] Ronny Kohavi and Barry Becker. 1996. UCI Machine Learning Repository.

[51] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[52] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[53] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analayzed the COMPAS Recidivism Algorithm. *ProPublica* 2016, 1 (2016), 1. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[54] Edward O Laumann, Peter V Marsden, and David Prensky. 1989. The boundary specification problem in network analysis. *Research methods in social network analysis* 61 (1989), 87.

[55] Elizabeth A Leicht and Mark EJ Newman. 2008. Community structure in directed networks. *Physical review letters* 100, 11 (2008), 118703.

[56] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* 22, 140 (1932), 55.

[57] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[58] Gianclaudio Malgieri. 2020. The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 154–166. https://doi.org/10.1145/3351095.3372868

[59] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

[60] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Virtual, 161–172.

[61] Julia Powles and Helen Nissenbaum. 2018. The Seductive Diversion of Solving Bias in Artificial Intelligence.

[62] Tim Räz. 2021. Group Fairness: Independence Revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 129–137. https://doi.org/10.1145/3442188.3445876

[63] Urša Reja, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar. 2003. Open-ended vs. close-ended questions in web questionnaires. *Developments in applied statistics* 19, 1 (2003), 159–177.

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.

2939778

[65] Tom Ron, Omer Ben-Porat, and Uri Shalit. 2021. Corporate Social Responsibility via Multi-Armed Bandits. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 26–40. https://doi.org/10.1145/3442188.3445868

[66] Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. Sage Publications, Los Angeles, CA, USA.

[67] Shalom H Schwartz and Wolfgang Bilsky. 1990. Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of personality and social psychology* 58, 5 (1990), 878.

[68] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[69] Richard A Shweder, Nancy C Much, Manamohan Mahapatra, and Lawrence Park. 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. *Morality and health* 119 (1997), 119–169.

[70] Heather L Stuckey. 2015. The second step in data analysis: Coding qualitative research data. *Journal of Social Health and Diabetes* 3, 01 (2015), 007–010.

[71] Vinith M. Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. 2021. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 723–734. https://doi.org/10.1145/3442188.3445934

[72] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, Las Vegas, Nevada, USA, 990–998.

[73] Renata Tesch. 2013. *Qualitative research: Analysis types and software*. Routledge, New York, NY, USA.

[74] Paul Whitney and Desiree Budd. 1996. Think-aloud protocols and the study of comprehension. *Discourse Processes* 21, 3 (1996), 341–351.

[75] Michael Williams and Tami Moser. 2019. The art of coding and thematic exploration in qualitative research. *International Management Review* 15, 1 (2019), 45–55.

[76] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[77] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 547–558. https://doi.org/10.1145/3351095.3375709

A ANALYZING THEMES VIA TOPIC MODELING

To discover and analyze themes across the corpus of FAccT publications, we perform topic modeling on the FAccT proceedings from 2018-2021. Our modeling complements authors' self-described keywords and CCS concepts; it enables us to take an unsupervised approach to elicit broader themes in FAccT's conferences. In the analysis that follows, we identify prevalent motifs and temporal patterns, some of which have thus-far remained elusive to the community, and which help provide an understanding of FAccT's incipient disciplinary identity. To elicit topics, we perform Latent Dirichlet Allocation (LDA) [9]. LDA enables us to model each paper as distribution over topics, where each topic representing a distribution over vocabulary words. In other words, at a high-level a topic can be viewed as a set of frequently-present, co-located words; LDA lets us find the topics to which a FAccT paper belongs, based on the individual words that the paper contains. More precisely, for a vocabulary $\mathcal{V}$ and a set of topics $\mathcal{T}$, if a particular word $w \in \mathcal{V}$ has a high probability of being in $t \in \mathcal{T}$, a paper that contains w has a higher probability of being about topic t.

Data curation. LDA computes a topic model using a bag-of-words representation of text. We therefore needed to prepare the FAccT proceedings, such that individually-isolated document tokens represented semantically-meaningful words. For example, we removed numerical results tables and math syntax, as individually tokenized numbers and symbols are divorced from their context-specific semantic meaning. We further describe our data curation process in the Jupyter notebook in our online repository. This process required a significant effort of code-assisted manual data cleaning, after which we manually verified that the cleaned FAccT corpus preserved the integrity of the original papers' text. We chose to include the 186 full-length archival proceedings papers to train our model, as a large portion of the non-archival abstracts were not papers, but rather tutorials and other in-conference community-based sessions and performances. We opted not to selectively arbitrate which contributions to include and exclude, and instead included only archival submissions.

Training procedure. The 186 papers in our dataset constitute a small-text corpus. We therefore looked to prior successful small-text topic analyses to inform our training procedure. Following recently published work [20], we use a Python wrapper of the MALLET library to train our model [5, 59]. MALLET, unlike the more-popular Python-based `gensim` library, uses Gibbs sampling [32] for the LDA algorithm's underlying sampling method. Gibbs sampling is an exact Markov chain Monte Carlo technique [15], which Cooper et al. [20] notes has better performance for small-text corpora than inexact, variational-inference LDA implementations [45]. For the documents submitted to LDA in training, we chunk the FAccT papers into contiguous segments of 200 words, following the intuition concerning hyperparameter tuning for Gibbs-sampling-based LDA described in the well-cited paper by Griffiths and Steyvers [35].

Hyperparameter selection.

Choosing the number of topics k for the model to learn requires some domain expertise and some degree of human intervention [17, 20, 35]. We began our experiments with $k = 20$, and tried larger and smaller k, guided by general advice in Griffiths and Steyvers [35], and the results reported reflect those that worked the best when confirming the output topics ($k = 22$). From performing this process for the 5 trained models ($k = 19, 20, 21, 22, 23$), we selected the model with $k = 22$ topics for our analysis. A selection of topic labels, with a subset of topic-specific words, can be found in Table 5.

We chose chunk sizes in such a way that ensured most document lengths are similar: documents are processed as a bag-of-words batch, so using a constant size serves to help normalize. We note that it does not help to chunk based on semantically meaningful sections because LDA operates on a bag-of-words representation, so we lose any semantic relationship between words (or higher level structure, like sections). When removing stop-words, we do not include the resulting chunk for training if it contains fewer than 20 words. In total, this procedure yields a 7149-document training corpus, composed using a vocabulary of 23764 words, with an average of 117.2 unique words per document.

Additional figures and results tables. We provide comprehensive results in our repository. These results include unnormalized results grouped by year, the dataframes used to generate the heatmaps in the paper, as well as by-paper topic distributions for each of the 186 papers in FAccT. We also include the full topic outs (i.e., the words for each topic).

B CITATION NETWORK ANALYSIS

Here, we describe in detail our data extraction process [B.1], the community detection algorithm we use [B.2], and the method we use to identify themes from communities [B.3].

B.1 Citation Data

For our citation analysis, we are interested in observing all articles appearing in FAccT proceedings and all papers that cite FAccT papers or are cited by FAccT papers. We use 2 datasets: the AMiner citation network dataset [72] and the Semantic Scholar Open Research Corupus (S2ORC) [57], to verify the robustness of our method. Each dataset has unique drawbacks: the AMiner data is incomplete in its resolution of citation links and does not give us complete lists of references to/from FAccT papers, while S2ORC had its latest release in 2020 and does not include FAccT 2021 papers. For the papers from 2018 to 2020, however, S2ORC resolves references better than AMiner.

B.1.1 AMiner Dataset: We use the 13th iteration of the AMiner dataset (released May 2021). Each entry in the dataset includes article metadata and contains fields for author list, venue, year of publication, and referenced articles. We obtain the subset of papers relevant for our analysis in three steps. We use the venue field of the dataset to filter in the papers published in FAccT. Since some papers have missing venue fields, as a second step, we manually search for missing papers using string search on words from paper titles or author names across the entire AMiner dataset. The first two steps give us a total of 207 FAccT papers. For the third step, we iterate over the entire dataset and extract articles that either cite, or are cited by, the articles in our seed set.

B.1.2 Semantic Scholar Open Research Corpus (S2ORC) Dataset: S2ORC is similar in metadata format to AMiner and we perform a similar 3-step procedure as described in the section on AMiner.

Table 5: A selection of results from our LDA-based topic model: A subset of topics, a corresponding sample of 4 words in the topic, and the (normalized) top-weighted paper for the topic.

Topic Label	Sample Words	Top Paper
`fairness/ optimization`	optimal fair cost constraint	Ron et al. [65]
`futility/ welfare`	group social utility welfare	Heidari and Kleinberg [40]
`policing`	crime allocation police policing	Akpinar et al. [4]
`law/rights`	legal law discrimination rights	Wilson et al. [76]
`image-classification`	images datasets vision face	Yang et al. [77]
`user-study`	participants human accuracy decision	Lai and Tan [52]
`health`	privacy health clinical patients	Suriyakumar et al. [71]
`fairness/ sensitive-attributes`	fairness protected sensitive attribute	Räz [62]
`fairness/ representation`	fairness representativeness moral representative	Malgieri [58]

Since S2ORC was released in 2020, we extract data and citation links corresponding only to the first three editions (2018-2020) of the conference.

B.2 Community Detection

For both of our datasets, we construct a directed citation network from the extracted set of papers (if an article A cites an article B, this is represented by a directed edge from A to B). We then identify communities within the resulting network. Here, we describe the algorithm we use to detect communities: a variation [23] of the Louvain algorithm [10] that incorporates edge direction in its optimization objective.

The Louvain algorithm and its variations rely on the idea of Modularity (Q), which is defined as the fraction of edges that fall within communities minus the expected fraction of such edges if they were distributed at random. More formally, modularity is given by the following expression:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{c_i, c_j} \tag{1}$$

Here, A is the adjacency matrix of the graph (A_{ij} is 1 if there exists an edge between i and j and 0 otherwise), m is the total number of edges in the network, k_i is the degree of vertex i, and δ is the Kronecker delta function, which takes the value 1 if both its arguments are equal and 0 otherwise.

An algorithm such as Louvain with modularity Q as its optimization objective disregards edge direction in a directed graph. It is, in fact, possible (and common practice) to disregard edge directions and to instead use the communities obtained from the corresponding undirected graphs, as is done by the Louvain algorithm. However, incorporating direction information can lead to the identification of more coherent communities as described by Leicht and Newman [55], who propose a formulation of *directed* modularity, that can be used with the Louvain algorithm to incorporate edge direction information while detecting communities.

The change in the optimization objective can be motivated as follows: traditional (undirected) modularity values are high when a *statistically surprising* fraction of edges in a network fall within the chosen communities. Leicht and Newman [55] extend this idea further to motivate their definition for directed modularity by suggesting that *any* statistically surprising configuration should contribute to an increase in modularity. Next, they consider two vertices—A and B—in a directed network such that A has a high out-degree and a low in-degree and B has a high in-degree and a low out-degree. In such a case, A → B edges are more likely than B → A edges. Therefore, B → A edges are more statistically surprising, and should contribute more towards modularity.

Intuitively, in the case of citation networks, A corresponds to an article with a small number of citations while B corresponds to an article with a larger number of citations. In this case, a B → A edge (B citing A) suggests more strongly that both A and B address the same area than an A → B edge does.

Leicht and Newman [55] formalize this by proposing a measure that can be used as an equivalent of modularity in directed graphs. This quantity, which we denote Q' is given by the following:

$$Q' = \frac{1}{m} \sum_{ij} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta(c_i, c_j) \tag{2}$$

Here, in addition to the symbols introduced in Equation 1, k_i^{in} and k_j^{out} are the in-degree and out-degree of vertices. This is the optimization objective we use while applying the Louvain method to perform community detection.

B.3 Identifying Themes & Naming Communities

To identify the areas or research directions each of the communities deals with, we study the titles of all papers present in a given community. Concretely, we compute unigram frequencies of terms appearing in paper titles and observe the top 10 terms to estimate the themes that the community deals with. We use both raw and weighted unigram-frequencies. We calculate top terms using two

methods: 1) using the raw unigram frequency of terms appearing in the titles and 2) weighting term frequencies by the in-degree of the paper they appear in.

We get very similar results from our analysis on both datasets. We present the additional results from the AMiner dataset in our GitHub repository.

C NSF IMPACT DEFINITIONS

In our questionnaire, we adopted the terms 'intellectual merit' and 'broader impact' from NSF (National Science Foundation), which defines:

- **Intellectual merit** as the contribution to advancement of knowledge and understanding. (Criteria include sound rationale and reasoning motivating the research, presenting creative, original, or potentially transformative concepts/approaches, well-organized execution of the research, positive scholarly impact within or outside the field).
- **Broader impacts** as benefits to society and contributions to the achievement of specific, desired societal outcomes. (Examples include empowering disadvantaged or marginalized individuals and communities; improving equity of access to opportunities; improving literacy and engagement of researchers, practitioners and the public).

D SURVEY DEVELOPMENT

In this section we briefly describe the development and validation process in designing, piloting, and deploying our questionnaire. The process unfolded over numerous stages of exploratory work, group feedback sessions, in-person interviews and think-aloud protocols, followed by our final iteration of view solicitation: a survey sent out to all FAccT authors with publicly available emails.

We began by drafting a set of relevant research questions and survey questions, which we presented to our research groups in two sessions. These sessions helped identify additional research inquiries, methods, sources of confusion, and relevant similar studies. After updating our questionnaire, we did two rounds of pilot surveys using think-aloud protocols [26, 27, 29, 74], where the two researchers each did (approximately) half-hour-long think-aloud procedures while subjects filled out the questionnaire, followed by half-hour-long interviews soliciting participants' views on the questionnaire. After conducting these four interviews, the questionnaire was updated to reduce ambiguities in how respondents interpret questions and control questionnaire length. The second round of pilots consisted of three hour-long interviews of the same nature, which were significantly more consistent and led to only minor language changes in the questionnaire.

Once finalized, the questionnaire was sent to every author of a FAccT conference publication, whose email was available either on the FAccT paper or on a personal website. In addition to attaching the questionnaire, we invited participants to participate in extended interviews. After receiving only three volunteers, however, we decided to forego interviews. We made this decision in order to protect the privacy of the research subjects and because we would not have been able to reach saturation with only three interviews.

E CHOOSING 5 FACCT TOPICS FROM TRACKS

We provide more granular details about how the five "topics" in our questionnaire (as seen in Figure 4) were chosen from the history of tracks in FAccT CFPs. We aggregated these tracks over all years, therefore, the topics chosen do not directly or exhaustively represent these tracks, but rather are winnowed from the full set. We note that the topics do not exactly match the "tracks" for two reasons: First, FAccT's tracks varied in terminology and subject year-over-year. Second, the pilot survey procedure, which consisted of 7 think-aloud interviews, uncovered certain tracks that led to inconsistency and confusion. In particular, the 'Security and Privacy' track was omitted from this analysis, because it significantly derailed survey responses during pilots. These participants became confused because they weren't aware of a security and privacy track.

F CODING OF FACCT PUBLICATIONS

Below we include a table representing the exhaustive set of codes used to analyze four years of FAccT papers. Two authors (Laufer and Heidari) were responsible for coding and used the papers' full text. Codes were not mutually exclusive; papers were assigned all relevant codes.

Table 6: Codes used to categorize and analyze FAccT papers.

Category	Code Type	Code
Title	Open-ended	e.g., "A Statistical Test for Probabilistic Fairness"
Synopsis	Open-ended	e.g., "Statistical hypothesis test for unfair classifiers"
Qualitative Research Design	Categorical	Scholarship review and critique Phenomenology Ethnography Case Study Grounded Theory Narrative Historical Action Discourse Analysis
Quantitative Research Design	Categorical	Experimental Survey Causal-comparative Correlational Cross-sectional Longitudinal
Field	Categorical	STEM Law Philosophy Social sciences and humanities
STEM Contributions	Categorical	Algorithm development Evaluations, metrics, measures Mathematical models and analysis Package, library, toolbox
Topic: Fairness	Categorical	Discrimination/group-level (un)fairness Individual-level (un)fairness Subgroup/intersectional (un)fairness Causal/counterfactual perspectives Tradeoffs Interventions and algorithms Types of biases Resource allocation/fair division
Topic: Transparency	Categorical	Transparency (audit, reproduce, data sharing, proprietary) Explainability (human-understandable translation) Interpretability (as inherent feature of model/algorithm)

Table 7: Codes used to categorize and analyze FAccT papers - Continued.

Category	Code Type	Codes
Topic: Accountability	Categorical	Human rights and freedoms, due process, recourse Policymaking, governence, and regulatory frameworks Professional codes, institutional procedures, industry standards Oversight and auditing mechanisms, compliance, liability Models from historically marginalized perspectives
Topic: Long-term/social impact	Categorical	Strategic behavior and its consequences Sequential decisions and interventions Feedback loops Polarization Trust/disinformation
Topic: Others	Categorical	Privacy, profiling, surveillance Human factors Other desiderata
Applications/Domains	Categorical	Internet Advertising Recommendation Systems E-commerce Social Media Entertainment and Media (Criminal) Justice System Law enforcement and policing Lending Healthcare/medical Hiring/employment Social services Design and Robotics Computer Vision Software
Off-the-shelf datasets	Categorical	COMPAS Crime and Communities Adult income German credit FICO HMDA MovieLens IMDB LSAT Student MNIST CIFAR-10
Original Datasets	Open-ended	e.g., "US EPA risk assessments"
Type of AI	Categorical	NLP Vision Deep Learning Systems (PL, DL, ...)
Misc Notes/Questions	Open-ended	

G CODING OF OPEN-ENDED SURVEY RESPONSES

Table 8: Coding and categorization of FAccT affiliates' opinions, criticisms, and suggestions for improvement.

Applicability	Category	Criticisms/Codes	Suggestions
Conference	Organization	Peer review quality	Separate track for HSA, STEM, and non-academic work Separate reviewing pools Reward truly interdisciplinary contributions Add a stage after acceptance for constructive criticisms Clarify standards for (interdisciplinary) contribution Make it more transparent Design education/onboarding process for reviewers
		Over-curation	Move toward more inclusive acceptance criteria
		Industry influence	Clarify conflict-of-interest policy for authors Implement funding disclosures Limit corporate involvement in conf. organization
		Misc.	Don't change CFP and deadlines Consider carbon footprint of conference
Conference	Relations	Industry	
		Public sector and governance	
		Other disciplinary venues	
Community	Culture	Activism-oriented vs. scientific agendas	Acknowledge the value of non-technical approaches Acknowledge the value of math models/solutions Clarify the academic, audit, & activist role of FAccT
		Lack of inclusivity	Promote CRAFT
		Non-constructive criticism	
		Ingroup-outgroup dynamics	
		Animosity toward STEM	
		Misc. (risk for junior researchers)	
Community	Interdisciplinarity	Collaborations	Emphasize translational research
		Communications	Develop shared conceptual infrastructure/vocabulary
		Contribution and novelty	Clarify standards for (interdisciplinary) contribution
Scholarship	Insularity	Lack of practical impact	Engage w/ "street-level bureaucrats" & practitioners
		Lack of diversity and inclusion	Outreach to communities (e.g., financial assistance)
		Intellectualism / echo-chamber dynamics	
		Solutionism	
		Lack of engagement with domain expertise	
		Lack of public and community engagement	
Scholarship	Narrow inquiry	Fairness	Broaden inquiry to notions of justice Address systemic oppression Include more STS contributions
		Machine learning	
		Western and US-centric values	
		Risk of "fair-washing"	
Scholarship	Quality & rigor	Ontology	Encourage conceptual work
		Lack of critical eval. of math assumptions	
		Discouraging mathematical contributions	

H QUESTIONNAIRE COPY

Introduction

This questionnaire aims to solicit your views on FAccT (Fairness, Accountability, and Transparency) scholarship, in particular, publications that have appeared in the ACM FAccT conference. Your answers will contribute to a research project studying research trends, gaps, and opportunities in the field.

The survey contains 3 main sections:
- Views on FAccT scholarship and recommendations for future improvements
- The intellectual merit and broader impact of several FAccT *research topics*
- The broader impact of FAccT scholarship in several *application domains*

The main questionnaire will be followed with an (optional) set of questions about your background and affiliation with the FAccT community.

To be mindful of your time, we have made all of the questions optional to respond to, but we would appreciate your input on as many of them as you believe you are qualified to respond to. Please do NOT respond to questions that you do not feel sufficiently informed/qualified to answer. At any point during the questionnaire, you can skip to the last section and submit the form (even if incomplete).

Views and Recommendations

This section seeks to elicit your opinion about several facets of FAccT scholarship, including
- Its most significant critiques
- Recommendations for future improvements

1. To what extent has the FAccT conference exposed its members to insights and ideas from other disciplines?

Mark only one oval.

	1	2	3	4	5	
Very low exposure	◯	◯	◯	◯	◯	Very high exposure

2. Reviewers of the FAccT conference come from various disciplines and backgrounds. In your experience, how has this impact the quality of the conference's peer review?

Mark only one oval.

	1	2	3	4	5	
Has severely impeded it	◯	◯	◯	◯	◯	Has significantly improved it

3. Many FAccT publications use publicly available datasets as part of their analysis. What is your overall assessment of the quali of these datasets for obtaining significant insights?

Mark only one oval.

	1	2	3	4	5	
Very Low Quality	◯	◯	◯	◯	◯	Very High Quality

4. Are there any moral or social values (sufficiently distinct from Fairness, Accountability, and Transparency) that you believe FAccT scholarship should address in near future?

5. What do you consider to be the most important criticisms of FAccT scholarship to date?

6. How do you believe the FAccT conference can address the above issues and limitations in the near future?

7. Any additional thoughts about FAccT?

Research Topics

In the questions that follow, you will be asked to rate the "intellectual merit" and "broader impact" of previous FAccT publications addressing several research topics, as well as how much FAccT scholarship should prioritize each topic.

The topics in this section are extracted from the "areas/tracks" of prior ACM FAccT conferences. (They are not necessarily mutually exclusive or exhaustive).

We have adopted the terms "intellectual merit" and "broader impact" from NSF (National Science Foundation), which defines:

- "Intellectual merit" as the contribution to advancement of knowledge and understanding. (Criteria include sound rationale and reasoning motivating the research, presenting creative, original, or potentially transformative concepts/approaches, well-organized execution of the research, positive scholarly impact within or outside the field).

- "Broader impacts" as benefits to society and contributions to the achievement of specific, desired societal outcomes. (Examples include empowering disadvantaged or marginalized individuals and communities; improving equity of access to opportunities; improving literacy and engagement of researchers, practitioners and the public).

Scoring Instructions

For each of the following topics, please score with 1 (indicating very low) up to 5 (indicating very high):

- The intellectual merit of *previous* FAccT publications on the topic.

- The broader impact of *previous* FAccT publications on the topic.

- Your assessment of whether *future* FAccT scholarship should prioritize this topic.

8. Algorithm Development
(e.g. fairness-enhancing algorithms; interpretable and explainable models)

Mark only one oval per row.

	1 (very low)	2	3	4	5 (very high)
Intellectual Merit					
Broader Impact					
Should FAccT Prioritize this Topic?					

9. Community-based Approaches and Human Factors

(e.g. participatory algorithm design; community designed and maintained systems; activism-driven technological change, human-computer interaction; humans-in-the-loop; information visualization; UX design)

Mark only one oval per row.

	1 (very low)	2	3	4	5 (very high)
Intellectual Merit					
Broader Impact					
Should FAccT Prioritize this Topic?					

10. Data and Algorithm Evaluation

(e.g. metrics; audits; data collection and curation)

Mark only one oval per row.

	1 (very low)	2	3	4	5 (very high)
Intellectual Merit					
Broader Impact					
Should FAccT Prioritize this Topic?					

11. Law, Policy, and Governance

(e.g. data protection; non-discrimination; fair procedures; human rights; oversight mechanisms; organizational governance; codes of ethics; models from historically marginalized perspectives)

Mark only one oval per row.

	1 (very low)	2	3	4	5 (very high)
Intellectual Merit					
Broader Impact					
Should FAccT Prioritize this Topic?					

12. Philosophy, Historical and Cultural Analysis

(e.g. philosophical foundations of machine learning; values in scientific inquiry; social epistemology; moral, legal and political philosophy of data and AI; Interrogating foundational concepts; Bridging critical concepts across fields)

Mark only one oval per row.

	1 (very low)	2	3	4	5 (very high)
Intellectual Merit					
Broader Impact					
Should FAccT Prioritize this Topic?					

13. Please briefly describe how you believe "impact" *should* be defined for FAccT scholarship.

Application Domains

The following section is concerned with several real-world, socially high-stakes domains in which algorithmic, data-driven tools have been utilized.

For each of the following domains, we would like you to score (with 1=very low up to 5=very high) the broader impact of previous FAccT publications for the domain's practices

A reminder that we have adopted the term "broader impact" from NSF (National Science Foundation), which defines "Broader impacts" as benefits to society and contributions to the achievement of specific, desired societal outcomes. (Examples include empowering disadvantaged or marginalized individuals and communities; improving equity of access to opportunities; improving literacy and engagement of researchers, practitioners and the public).

14. FAccT's Broader Impact on Domains

Mark only one oval per row.

	Unfamiliar	1 (very low)	2	3	4	5 (very high)
Criminal Justice System						
Education						
Healthcare/Medical						
Hiring/Employment						
Internet Advertising						
Social Media						
Social Services						

Information About You

Please tell us about yourself to the extent that you are comfortable. Remember that the questionnaire is anonymous. (All categorical alternatives have been ordered alphabetically).

15. Which of the following areas of scholarship best describes your expertise?

Check all that apply.

- [ ] HSA (Humanities, Social Sciences and the Arts)
- [ ] STEM (Science, Technology, Engineering, Mathematics)

Other: [ ] _____

16. If you have played any formal role in the FAccT conference (including the main conference, as well as its workshops, tutorial CRAFT sessions), select all that apply.

Check all that apply.

- [ ] Attendee
- [ ] Author
- [ ] Organizing Committee Member
- [ ] Reviewer

Other: [ ] _____

17. How many years have you been affiliated with the FAccT conference?

Mark only one oval.

	0	1	2	3	4	5	
0 year of involvement	◯	◯	◯	◯	◯	◯	5 years of involvement

18. Which of the following best describes your political views? (Please pick the closest approximation if none is exactly right).

Mark only one oval.

- ◯ Conservative
- ◯ Liberal
- ◯ Libertarian
- ◯ Other: _____

19. Do you belong to a marginalized/disadvantaged group or community?

Mark only one oval.

- ◯ No
- ◯ Yes

20. If you would like to communicate anything else about yourself with us, please do so here.

21. Thank you so much for your participation. If you have any feedback for us about the questionnaire, please leave your comments here, or email the PI.

Interested in participating in one-on-one, open-ended interviews?

Please contact the PI (hheidari@cmu.edu) if you are interested in participating in a longer interview with the research team to provide us with your detailed thoughts assessments of FAccT.

This content is neither created nor endorsed by Google.

Google Forms

I CONSENT FORM FOR SURVEY RESPONDENTS

Carnegie Mellon University

Online Consent Form

This study is part of a research study conducted by Professor Hoda Heidari at Carnegie Mellon University.

Summary: The goal of the proposed study is to understand research trends, gaps, and opportunities in the growing field of Fairness, Accountability, and Transparency (FAccT) by surveying a subset of research community members.

Purpose: The purpose of the research is to elicit your opinion about the FAccT scholarship to date. Fairness, Accountability, and Transparency (FAccT) for Artificial Intelligence (AI) is a thriving area of research addressing several key societal and ethical considerations around the growing use of AI in society. An ACM conference bearing the same name has been the central venue for scholars from various disciplines to come together, provide peer feedback to one another, and publish their work. The goal of this study is to understand this field of research through the lens of its community members. In particular, we aim to elicit members' opinions about the field's progress so far, impactful areas for future research efforts, and the nature of interdisciplinary work in this area.

Procedures: You (the study participant) will answer a set of questions about the following aspects of the FAccT Scholarship:

- The impact and urgency of central research topics
- The impact of the scholarship on real-world applications
- Interdisciplinary awareness and collaborations
- Moral and social values that FAccT must address
- Major critiques of the literature to date
- Suggestions to address these issues in the near future

In addition to the above questions, we will ask you to provide us with additional information about your background and affiliation with FAccT. The purpose of these questions is for us to detect any significant variations in responses across the corresponding dimensions. Answering this part is entirely optional. We expect that the survey will take **less than 15 minutes** to complete.

All your responses will be saved in an anonymized manner, so we will not be able to map your responses to your identity. If you would like us to acknowledge your participation in any future reports on the study, or associate your name with your responses, please email the Principal Investigator at hheidari@andrew.cmu.edu.

Participant Requirements: Participation in this study is limited to individuals age 18 and older.

Compensation & Costs: You will not receive monetary compensation for participating in this study. There will be no monetary cost to you either.

Risks: The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities.

Version 1.2018

Carnegie Mellon University

Benefits: There may be no personal benefit from your participation in the study.

Future Use of Information: Once we remove all identifiable information from your responses, we may use the data for our future research studies, or we may distribute the data to other researchers for reproducibility purposes. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that preserves your anonymity.

Confidentiality: By participating in this research, you understand and agree that Carnegie Mellon may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your consent form will be stored in a secure location on Carnegie Mellon property and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by Carnegie Mellon and published and/or disclosed by Carnegie Mellon to others outside of Carnegie Mellon. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by Carnegie Mellon. Note that per regulation all research data must be kept for a minimum of 3 years.

Right to Ask Questions & Contact Information: If you have any questions about this study, need additional information, or wish to withdraw your participation, you can contact the Principal Investigator at hheidari@andrew.cmu.edu.
If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the Office of Research Integrity and Compliance at Carnegie Mellon University. Email: irb-review@andrew.cmu.edu . Phone: 412-268-1901 or 412-268-5460.

Voluntary Participation: Your participation in this research is voluntary. You may discontinue participation at any time during the research activity. You may print a copy of this consent form for your records.

By proceeding to respond to the questionnaire, you acknowledge that you are 18 or older; you have read and understood the information above, and you are willing to participate in this research.

Version 1.2018