Are AI Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences?

Daniel E. Acuna
School of Information Studies
Syracuse University
Syracuse, NY 13244, USA

Lizhen Liang
School of Information Studies
Syracuse University
Syracuse, NY 13244, USA

ABSTRACT

Even though computer science (CS) has had a historical lack of gender and race representation, its AI research affects everybody eventually. Being partially rooted in CS conferences, “AI ethics” (AIE) conferences such as FAccT and AIES have quickly become distinct venues where AI’s societal implications are discussed and solutions proposed. However, it is largely unknown if these conferences improve upon the historical representational issues of traditional CS venues. In this work, we explore AIE conferences’ evolution and compare them across demographic characteristics, publication content, and citation patterns. We find that AIE conferences have increased their internal topical diversity and impact on other CS conferences. Importantly, AIE conferences are highly differentiable, covering topics not represented in other venues. However, and perhaps contrary to the field’s aspirations, white authors are more common while seniority and black researchers are represented similarly to CS venues. Our results suggest that AIE conferences could increase efforts to attract more diverse authors, especially considering their sizable roots in CS.

CCS CONCEPTS

• **Computing methodologies~Artificial intelligence** • Social and professional topics~Computing organizations • Social and professional topics~Race and ethnicity • Social and professional topics~Gender

KEYWORDS

Artificial Intelligence; Ethics Conferences; Content and Citation Analyses; Science of Science

ACM Reference format:

Daniel E. Acuna & Lizhen Liang. 2021. Are AI Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences?. In *Proceedings of 2021 AAAI/ACM Conference on AI, Ethics and Society (AIES’21)*, May 19–21, 2021, Virtual Event, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3461702.3462616>

This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES '21, May 19–21, 2021, Virtual Event, USA.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8473-5/21/05.
<https://doi.org/10.1145/3461702.3462616>

1 Introduction

A great deal of research in artificial intelligence occurs in the context of computer science conferences. For example, the pioneering Neural Information Processing Systems (NeurIPS) conference served as a conduit for early collaborations between neuroscientists and research in brain-inspired techniques used in modern deep neural network research [1]. Computer science has naturally dominated research in this area: considerable improvements to AI have come from better algorithms, data structures, and hardware [2]. Recent cases of biases in AI systems have shaken the community and society, and made these CS conferences self-reflect on the social implications of their work [3]. This introspection has motivated the community to create new conferences that address these issues. Some of these efforts have resulted in guidelines and recommendations with policy implications [4]–[6]. However, it stands to reason to ask whether these new “AI ethics” (AIE) venues have different representations from traditional CS conferences: real-world biased AI decisions mostly affect communities already underrepresented in CS. Are we, as a field, replicating past issues with CS conferences in AIE conferences? How are the authors, institutions, countries, contents, and citations different? Are researchers affected by biases in AI better represented in these new venues? Here, we use meta-science [7] to start answering some of these questions.

The study of ethics in artificial intelligence has grown in prominence over the last several years. Classically, the concern started in distorted perceptions of the power of robots [8]. More recently, however, the widespread use of AI software has produced palpable real-world consequences. Cases of biases in job candidate screening [9], decisions in the justice system [10], and financial systems [11] have highlighted how important and pervasive these issues are. Public uproar and research communities have prompted the CS and AI fields to create new specialized conference venues where these issues can be addressed [12]–[14]. Two of the most prominent such conferences are the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) and AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), both started in 2018. FAccT and AIES have already produced theoretical frameworks and concrete software solutions to detecting biases [15] and fixing them [16]. While other conferences discuss AI ethics (e.g., 4S, EASST, CEPE, IACAP, SPT), FAccT and AIES tend to have stronger CS roots. The study of ethics in AI thus has enthusiastic venues where issues are starting to be addressed.

According to theories of liberalism and social justice, the legitimacy of a system is linked to the idea that people who are subject to the system should agree to it [17], [18] (but see [19] for contrasting views). Translated into AI ethics, it could be argued that AI systems should seek validation from those most affected by them—the underrepresented and unprivileged. The puzzle arises, however, when we consider that computer science has historically suffered from major issues related to sexism, racism, and lack of inclusion [20], and it is often perceived as a “masculine” and “white” discipline [21]. If issues of bias in AI are to be addressed in meaningful ways, liberalism and social justice suggest that we should strive to hear the opinions and research of those most affected by the technology. To the best of our knowledge, it is largely unknown whether these new AI ethics (AIE) conferences have better representations from these communities compared to the CS conferences associated with AI research. Not understanding if and how these communities are heard risks replicating and maintaining the faults that traditional CS conferences have suffered in the past.

Explicitly studying how a scientific field evolves has many benefits beyond the rhetorical and theoretical realms. From a *science of science* perspective [7], scientists and fields can sometimes get trapped in “fads” controlled by a few prominent and influential authors [22]. Therefore, it is desirable to understand the makeup of a field to avoid these issues. By looking at temporal trends, we can also predict future thematic foci [23]. Scientific bodies, journal editors, and conference general chairs can take this information and try to steer the field away from dead ends towards new challenges. For AI ethics research, there are other societal factors that come into play. AI ethics venues might actively invite researchers or the general public to be involved in discussions because they might understand historical failures of traditional CS communities [24]. Analyzing the characteristics and evaluation of who publishes is therefore essential.

In this work, we study and contrast the characteristics of authors, institutions, fields, countries, and citations of AIE and CS conferences. We use a large dataset of bibliometric data and analyze the top entities involved in both types of meetings. We study the temporal trends in field and country diversity and authors’ characteristics, including gender, race, and seniority. Finally, we explore how AIE conferences are different from CS ones by using the conference venues’ predictability as a function of their content. We study the keywords and tokens that are most predictive of work published in AIE and conclude that it is highly unique and discernible.

2 Research on bias in AI: a brief introduction

Artificial intelligence (AI) has attracted attention from academia and industry by showing promising performance in various tasks traditionally done by humans [25]. Companies and governments have been using systems powered by AI to help with tedious or labor-intensive tasks, including job candidate screening [9] and credit scoring [11]. These apparently boundless opportunities have not occurred without controversies.

Biases and other social issues of AI are not new. Even before recent artificial intelligence breakthroughs, computer programs have been found guilty of making biased decisions [10]. For example, simple expert systems for supporting health decisions—highly inspectable by their simple nature—have shown biases [26]. In the past, the benefits of expert systems capturing human experts’ decisions seemed to outweigh these potential shortcomings. However, the relatively low accuracy and high cost did not make them feasible at large scales [27]. While AI ethics issues have always lingered behind developers’ creations, their limited applicability has not been a cause for concern.

Recent breakthroughs in AI have made it so that the accuracy and complexity of systems have skyrocketed. These advances are especially true with the advent of deep learning [28]. These systems have also become viable because of new architectures, types of neurons, loss functions, and optimization methods [29]. Recent advances in Graphics Processing Units (GPUs) have made the training of these networks viable. More importantly, modern deep learning models used in many societal settings are many orders of magnitude more complex and non-linear than before. This complexity makes them challenging or impossible to interpret. Therefore, biases and discriminatory decisions are sometimes found once these systems are deployed en masse [3].

As AI was becoming increasingly common and impactful to society, AAAI/ACM Artificial Intelligence, Ethics, and Society was announced in 2017 to provide a platform for addressing these issues. ACM Conference on Fairness, Accountability, and Transparency (now called ACM FAccT) was announced in 2017 for the same purpose [12]–[14]. Because computer scientists are the ones who mostly develop AI systems, these conferences have a strong foundation in this discipline.

2.1 Conferences as an important source for the birth and evolution of ideas

New conferences targeting a specific topic stimulate conversations between researchers from different disciplines with a shared goal. Unlike typical computational research, knowledge from a wide range of fields, including computer science, mathematics, sociology, and public policy, is involved in addressing AI ethics issues. There are few places for researchers to join discussions about these issues than AI ethics conferences.

Similarly, there are few better places to study a field than conferences focused on that field. Through a conference, we can learn the field’s dynamics by studying the institutions, authors, and areas involved [30]. With temporal publication trends, we may understand how the dynamics has been changing and how the field is evolving [31]. By studying the publication and citation information of a conference and comparing it with other meetings, we may gain insight into the relationship between them and identify significant cross-disciplinary collaboration opportunities [32].

3 Using publications, institutions, content, and citations to analyze AI ethics conferences

In this work, we study the characteristics of AIE conferences and contrast them to popular traditional CS conferences. We do so from the science of science point of view [7], which uses a variety of quantitative methods to study scientific processes and research behaviors using publication data, citation data, and author and affiliation statistics. We now describe the data and methods used in our study.

3.1 Data.

We first need to identify publications and citations. We use the Microsoft Academic Graph (MAG) [33], which contains exhaustive publication, citation, authorship, and affiliation data. We use Semantic Scholar for content analysis, which includes the abstract of publications [34]. With MAG and Semantic Scholar, we can locate publications from our AI ethics and computer science conferences of interest.

For AIE conferences, we select FAccT and AIES for our analysis. For CS conferences, we choose the top 10 meetings by their combined impact and productivity. These CS conferences are AAAI, ACL, CVPR, ECCV, EMNLP, HLT-NAACL, ICCV, ICML, NAACL, and NeuIPS. These are a combination of NLP, Computer Vision, and general Machine Learning conferences, and likely deal with ethical issues. From both data sets, we are able to find 381 publications from AIES and FAccT conferences from 2018 to 2020. To make comparisons relevant, we restricted publications from CS conferences to the same time frame. We found 14,179 publications from our target CS conferences, published between 2018 and 2020.

Finally, we get geographical information about affiliations and author features related to citations such as h-index and fields of study from the data set. We retrieve features related to institutions and authors with publications accepted by those conferences from the affiliations.

3.2 Methods.

We are able to locate conferences from AI ethics conferences and popular traditional CS conferences by searching the normalized conference names from MAG and Semantic Scholar. With DOIs corresponding to articles published in conferences, we are able to get publication information, the author’s information, and affiliation information from MAG.

Gender and race estimation. In order to investigate the trend of gender and ethnic diversity of authors, we built a BERT-based prediction model for gender and race based on the author’s name. We used the Genni + Ethnea data set from [35], which is a large dataset that includes full names from a wide range of origins from around the world. The dataset contains pseudo-labels which are predictions from an ensemble of tools. We include the data set as part of the training and validation data set in order to ensure the diversity of names so that the model has better generalizability. For gender prediction, we combined prediction results from Genni, SexMac, and SSNgender using the majority vote rule. For

validation, we have included another data set released by the Social Security Administration which includes popular newborn names and their gender [36]. We aggregated gender categories to only “female”, “male” and “unknown”. For ethnicity prediction, we aggregated prediction results from Ethnea, which includes a wide range of 26 kinds of ethnicities. For validation, we use a dataset with name and ethnicity information from Wikipedia created by [37]. For the combined data set, we map the original ethnicity labels into “Asian”, “Hispanic”, “Black” and “White”. We combine the first name and the last name from the data set and generate both character-level tokens and word-level tokens for predictions. We estimated the performance of our model by cross-validating on our training datasets (“Val” in Table 1 and Table 2) and by validating on external datasets (“SSA” in Table 1 and “Wiki” in Table 2). We estimated the F1 score, Accuracy, and area under the precision–recall curve. There are other popular automated methods to predict gender and ethnicity of names, for example, genderize [38], gender-guesser [39], and gender API [40]. The performance of our method is similar to what has been reported in the literature before (e.g., see Table 4 in [41])

	Male	Female	Unknown
F1 (Val)	0.961	0.975	0.889
Accuracy (Val)	0.972	0.979	0.862
AUC (Val)	0.993	0.996	0.966
F1 (SSA)	0.813	0.915	0.504
Accuracy (SSA)	0.711	0.885	0.664
AUC (SSA)	0.954	0.965	0.860

Table 1. Gender prediction performance

	Black	Hispanic	White	Asian
F1 (Val)	0.976	0.936	0.907	0.941
Accuracy (Val)	0.999	0.928	0.902	0.931
AUC (Val)	0.999	0.990	0.983	0.989
F1 (Wiki)	0.987	0.822	0.850	0.859
Accuracy (Wiki)	0.999	0.788	0.856	0.843
AUC (Wiki)	0.996	0.964	0.963	0.962

Table 2. Race prediction performance

Author’s impact and productivity (seniority). By analyzing the citation network in MAG, we calculate the h-index for each author who published in AIE or CS conferences in any given year. The h-index, or Hirsch index, is a measure of a researcher’s productivity and impact. For an author with an h-index of x , the author has received x or more citations for at least x of his or her publications [42]. We compared authors who have published in AIE conferences and authors who have published in CS conferences by comparing the average h-index and its changes over time.

Inter-conference impact. With the citation network, we could locate publications in CS conferences citing publications in AIE conferences and vice versa. For CS conferences, we were able to get the percentage of citations to publications from AIE conferences, and for AIE conferences, we were able to get the rate of citations to publications from CS conferences. By analyzing the cross-citation between the two kinds of conferences, we can discover how one field affects the other.

Publication diversity. We are also able to find each author’s publication in the last ten years and the corresponding field of each of those publications. For each author who has publications accepted by AIE conferences and CS conferences, we estimate an author’s field by using the most common field. With the estimated areas of authors, we are able to measure the diversity of a conference by calculating the entropy given by the author field distribution using

$$\text{entropy} = \sum_{i=1}^n P_i \times \log(P_i),$$

where n is the number of fields, and P_i is the frequency of that field in a conference. A high entropy indicates that a conference has high diversity while a low entropy indicates a low diversity. With the definition, we were able to measure the diversity of each conference in terms of fields.

Geographic diversity. We are able to reverse geocode the latitude and longitude of affiliation provided by MAG. We use the Python package `reverse_geocoder` [43]. We are able to identify the nationalities of institutions having publications accepted by AIE or CS conferences. With the method mentioned above, we are able to calculate the entropy of countries for each conference and measure each conference’s diversity in terms of nationality.

Conference content analysis. To further understand the differences between AIE and CS, we measure the differences between publications from different fields by measuring the use of words and topics. We use the abstracts provided by Semantic Scholar for such analysis. After cleaning the abstract corpora, we computed the term frequency-inverse document frequency (tf-idf) of tokens and trained a multinomial logistic regression to find predictive tokens for each conference. The cleaning process includes lowercasing, stemming, and stop-word removal. We used Porter stemmer to transform each word to its stem form [44]. With the stop-word removal function in scikit-learn, we removed frequent words (e.g., “the”). We then stratified the same number of AIE publications and publications from each selected CS conference for all our analyses.

4 Results

In this article, we are trying to understand the characteristics of artificial intelligence ethics (AIE) and computer science (CS) related conferences. We will explore the basic characteristics of each conference, temporal dynamics and differences between them, citation patterns, and the distinguishability of their documents.

4.1 Basic characteristics of conferences

We first examine both conferences’ basic characteristics using the most popular terms of the countries, institutions, and topics (see methods). The ranking is determined by the number of articles published by a country. We found that the top countries publishing in these venues are relatively similar (Table 3).

#	AIE conferences		CS conferences	
1	USA	74%	USA	42%
2	United Kingdom	9%	China	19%
3	Canada	2%	United Kingdom	4%
4	Netherlands	2%	Germany	3%
5	Germany	2%	Australia	2%
6	Brazil	1%	Canada	2%
7	Switzerland	1%	South Korea	2%
8	Norway	1%	Hong Kong	2%
9	New Zealand	<1%	Japan	2%
10	Spain	<1%	Switzerland	1%
	Others	6.21%	Others	18.19%

Table 3. Top 10 countries (regions) publishing from 2018 - 2020

However, while China makes up 19% of the authorships in CS conferences and is in the top 2 positions after the USA, it does not appear in the top 10 countries in AIE conferences. There are only eight authorships affiliated with China from 2018 to 2020 in AIES. In terms of institutions, we found that Google and CMU are at the top of both types of venues (Table 4). However, companies tend to be more involved in AIE conferences, appearing at the top 1 (Google), top 3 (IBM), and top 7 (Microsoft). Facebook has a negligible presence in AIE conferences with only four publications. Chinese companies are also involved but only in CS conferences (e.g., top 11: Tencent (not shown)).

#	AIE conferences		CS conferences	
1	Google	5%	Google	3%
2	Carnegie Mellon University	4%	Carnegie Mellon University	2%
3	IBM	4%	Chinese Academy of Sciences	2%
4	Stanford University	3%	Tsinghua University	2%
5	University of Oxford	2%	Microsoft	2%
6	Cornell University	2%	MIT	1%
7	Microsoft	2%	Stanford University	1%
8	Duke University	2%	Peking University	1%
9	University of Cambridge	2%	IBM	1%
10	UC, Berkeley	2%	UC, Berkeley	1%
	Others	65%	Others	80%

Table 4. Top 10 institutions publishing from 2018 - 2020

Finally, Table 5 shows the top fields of publication. Expectedly, computer science and artificial intelligence are at the top. However, in AIE conferences, medicine, economics, psychology, and sociology made it higher in the ranking, suggesting broader applicability of the ideas presented. These rankings show that there are differences in representability between these types of conferences.

#	AIE conferences	CS conferences
1	computer science 46%	artificial intelligence 52%
2	artificial intelligence 24%	computer science 36%
3	mathematics 2%	mathematics 2%
4	medicine 1%	mathematical optimization 1%
5	economics 1%	algorithm <1%
6	social media 1%	biology <1%
7	psychology 1%	adversarial system <1%
8	sociology 1%	medicine <1%
9	biology 1%	discrete mathematics <1%
10	mathematical optimization 1%	psychology <1%
	Others 5%	Others 17%

Table 5. Top 10 fields publishing from 2018 - 2020

4.2 Temporal trends and differences between AIE and CS conferences

Even though the AIE conferences analyzed started only in 2018, we can still attempt to understand temporal trends. We first wanted to examine the differences among demographic factors that are considered important for both conferences. One group of factors is related to the articles and institutions in those articles. Another group of factors is related to the authors themselves.

One important factor in conferences is to understand how diverse the fields presented are. Some conferences might prefer to be more focused and present a small number of fields while others might be less focused and cover a wider base. To quantify this “diversity,” we use entropy (see methods for a definition). In particular, a high entropy means that the group of authors represents a broad set of fields (Fig. 1a). We measure field diversity using this entropy across AIE and CS conferences from 2018 and forward (Figure 1a). Field diversity in AIE conferences grew significantly from 2018 to 2019 (two-sided bootstrap test [TSBT], $p < 0.01$) and also grew from 2019 to 2020 (TSBT, $p = 0.07$). The field diversity in CS conferences grew significantly from 2018 to 2019 (TSBT, $p < 0.0001$) but then it dropped significantly from 2019 to 2020 (TSBT, $p < 0.0001$). Across years, AIE conferences have significantly higher field diversity than CS conferences (AIE: $M=1.88$, $SE=0.5$; CS: $M=1.38$, $SE=0.10$, $p < 0.0001$). The changes in field diversity in AIE might represent changes across CS rather than something specific about AIE. To test this hypothesis, we performed a difference-in-difference analysis [45], where we compare the changes in field diversity from 2018 to 2020 in AIE and CS. Our analysis showed a significant difference between these changes (z-score: 2.44, $p < 0.02$). This result suggests that AIE conferences have higher field diversity than CS conferences; they have been increasing this diversity differently from how CS has been changing its field diversity.

Another important factor for conferences is to reach wide authorship of countries. Especially for conferences such as AIE, it might be desirable to have work presented from an equal distribution of institutions and countries. Similarly to how we measure field diversity before, we measure the country diversity

(Fig. 1b). We measure this diversity for the AIE and CS conferences from 2018 forward. We found that both AIE and CS conferences had a significant decrease in country diversity (AIE, $p < 0.001$; CS, $p < 0.001$). Moreover, we found that AIE conferences had significantly lower country diversity compared to CS conferences (z-score: -15.25, $p < 0.001$). We also found that AIE conferences had a decrease in diversity that was not significantly different from that of CS conferences ($p = 0.97$). These results suggest that AIE and CS conferences have decreased country diversity alike.

Figure 1. Temporal characteristics of AI Ethics (AIE) and CS conferences with S.E. error bars. (a) and (b) are entropies, which represent how nonuniform the distributions are. Higher values mean more “diversity.” (c) and (d) represent the proportions of male and white authors. (e) The h-index represents the “seniority” (impact + productivity) of researchers.

One important source of contention for CS-related conferences is its historical lack of gender and race diversity [20], [21]. A percentage of 67.6% of the authors in AIE conferences were estimated to be male ($SE=2.63\%$), while 78.8% of authors in CS conferences were estimated to be male ($SE=1.44\%$, Figure 1c). Both types of conferences saw a large decrease in male authorship from 2018 to 2020 (AIE=-12%; CS=-9%), and this decrease was not

significantly different among the conference types ($p = 0.568$), suggesting an overall increase in the number of non-male authorship in these venues.

A more equal race composition of authors should be desired in AIE conferences. We found that CS conferences have more authors of other races in recent years (white: $M=37.2\%$, $SE=1.1\%$. Asian: $M=48.6\%$, $SE=1.6$, $z\text{-score} = -2.23$, $p < 0.05$, 2018-2020). We found also that, on average, 52% of the authors in AIE conferences are white ($SE=5.5\%$) and there is a significant increase in white authorship from 2018 to 2020 (TSBT, $p < 0.05$, Figure 1d). We also found that there are no significant differences in black authorship between CS conferences and AIE conferences (AIE black authorship: $M=3.4\%$, $SE=0.0053\%$, CS black authorship: $M=2.7\%$, $SE=0.063\%$, $z\text{-score}: 0.46$, $p=0.646$). Taken together, these results suggest that AIE conferences have more white authorship than CS conferences, and both conferences have similar representation of black authors. Perhaps one of the AIE conferences' goals is to attract the next generation of (more junior) scientists who can change the culture in CS conferences. We tested whether this change is true by measuring the seniority of authorship using the h-index (see methods). We found that on average, both conferences have similar h-indices (AIE: $M=13.33$, $SE=0.96$; CS: $M=13.76$, $SE=0.55$), and these differences are not significant (Fig. 1e). While we found that CS conferences have seen significantly more junior authors (i.e., lower h-indices) from 2018 to 2020 (CS: h-index 2018: 14.62 ($SE=0.15$), h-index-2018: 11.43 ($SE=0.93$), $p < 0.0001$), the difference is not significant from AIE conferences (AIE: h-index 2018: 13.04 ($SE=1.08$); h-index 2020: 11.84 ($SE=0.92$), $p = 0.1328$). Both conferences have seen an average decrease in h-index of 0.41 points that is not significantly different across conferences ($z\text{-score}: -0.30$, $p = 0.757$). In conclusion, the makeup of both types of conferences in terms of seniority is relatively similar.

4.3 Gender composition of teams publishing

Gender composition of teams of co-authors shows how diverse or polarized the community is. We analyzed the gender composition of all teams of co-authorships published in either CS conferences or AIE conferences. On average, 32% of the teams publishing in AIE conferences are man-only and 9% of the teams published to AIE conferences are women-only. In CS, 31% of the teams published are man-only while only 2.5% are women-only. Such results show that AIE conferences have accepted more publications by women-only teams and fewer of man-only teams.

We then wanted to analyze whether team composition is different while controlling for the fact that single-gender teams are harder to produce for bigger teams. This is particularly important when comparing AIE and CS conferences as CS conferences tend to have significantly larger teams. We do this analysis by performing a fixed effect model that relates the conference type of a publication with the chance of having an all-male co-authorship team while controlling the base rate as a function of team size as a random factor. Indeed, we found that the fixed effect odds-ratio of the CS conference is -0.3577, which indicates an approximately 4.4% less chance of having an all-man team in CS compared to AIES. This

difference, however, is barely insignificant (odds-ratio=-0.3577, $z: -1.948$, $p=0.051$).

4.4 Citation patterns across AIE and CS conferences

Citations are an important part of understanding how publications, authors, and fields affect one another. We analyze these patterns between these two kinds of conferences. First, to understand whether AIE conferences are having an impact on CS conferences, we performed a regression analysis to estimate the effect of year on the number of citations from AIE to CS papers. We found a positive effect, ($t(18)=2.09$, $p = 0.051$), suggesting an increase of AIE impact on CS. However, we found that AIE conference papers have reduced their citation to CS conference papers. Using a regression model analysis, we found a non-significant negative association between year and citation from AIE to CS ($t(3) = -0.684$, $p = 0.543$). These results suggest that CS conferences are increasingly citing AIE conferences while AIE conferences rely less and less on CS conferences.

We then tested the hypothesis that people who publish CS conferences use AIE conferences as venues to publish something different than simply another CS article. We tested this hypothesis by evaluating how many people who publish in CS conferences also publish in AIE conferences. We found that only 0.92% of authors do this. However, we found that people who publish in AIE also publish in CS 25% of the time. This suggests that AIE conferences provide a venue for CS researchers to publish different work. Still, the fraction of scientists who publish at both conferences is very small (0.92%). This suggests that AIE is not taking people "away" from CS conferences, but rather, these conferences complement each other.

4.5 Discernibility of conferences

One of the goals of AIE conferences is to provide a venue for CS researchers to publish work exploring the effects of AI on society. However, one question is whether the work presented in AIE conferences is truly different from work presented in other CS conferences. To answer this question, we use the abstract of publications as a signal to estimate how well we can tell apart conferences, both AIE and CS. We first examine whether we can predict AIE vs. non-AIE conference based on an abstract (see methods for preprocessing) using a simple regularized logistic regression. This model had an accuracy of 96% with a precision of 0.9, recall of 0.52, and F1 score of 0.66. Predicting CS conferences has higher precisions and recalls ($P=0.95$, $R=0.99$, $F1=0.94$). This suggests that AIE and CS conferences have distinct features and serve different purposes.

Sometimes it would be useful to understand how much we can drill down on our predictive measures. For example, can we tell apart the precise CS conference based on the abstract while separating them from AIE conferences? We performed a multinomial logistic regression model with regularization to answer this question. The model took as input the tf-idf vectors of abstracts (see methods).

	AIE	AAAI	ACL	CVPR	ECCV	EMNLP	HLT-NAACL	ICCV	ICML	NAACL	NeurIPS
AIE	96%	3%	0%	0%	0%	0%	0%	0%	1%	0%	0%
AAAI	14%	35%	5%	0%	7%	4%	2%	1%	13%	0%	18%
ACL	3%	4%	33%	1%	0%	7%	26%	0%	1%	24%	1%
CVPR	0%	1%	0%	36%	28%	1%	0%	24%	5%	1%	5%
ECCV	1%	0%	1%	23%	30%	0%	0%	35%	4%	1%	6%
EMNLP	0%	12%	14%	1%	0%	27%	18%	0%	4%	22%	1%
HLT-NAACL	1%	0%	26%	0%	0%	6%	50%	0%	7%	10%	0%
ICCV	0%	0%	1%	27%	20%	0%	0%	34%	8%	2%	7%
ICML	4%	4%	2%	4%	2%	2%	2%	2%	57%	1%	19%
NAACL	3%	1%	8%	0%	0%	17%	5%	0%	0%	63%	3%
NeurIPS	0%	0%	0%	0%	0%	3%	0%	0%	3%	0%	94%

Table 6. Confusion matrix. Predictability of conferences based on paper abstracts. AIE: AI Ethics conferences (articles from 2018-present). Rows are the true labels.

There are some clear patterns in the confusion matrix (Table 6). As shown before, AIE conferences are relatively straightforward to tell apart from all other conferences, with only 3% of the errors going to the AAAI conference. There is clear overlap and difficulty in telling computer-vision related conferences, such as CVPR, ECCV, and ICCV, as shown by the shared error rates among them. A similar effect happens among the Natural Language Processing (NLP) related conferences: ACL, EMNLP, HLT-NAACL, and NAACL. Finally, the more theoretically-inclined conferences (ICML and NeurIPS) have some overlap. However, the NeurIPS conference abstracts are easier to classify than ICML abstracts. These results suggest that abstracts can allow us to distinguish between conferences in a precise manner.

It would be useful to understand *why* we can tell apart the conferences, especially AIE vs. the other CS conferences. We can investigate this question by looking at the words that more prominently affect the multinomial logistic regression scoring. In Table 7, we display the top 10 words, ordered by coefficient weight, predictive of the class. This list displays not just the most important words of the particular class but also the most important words with respect to all other classes being considered in the prediction. We can see that the words that distinguish AIE conferences are precisely the topics that these conferences are set to cover. Namely, they cover fairness, norms, social, and ethical aspects. We can also see the word explainability and bias (words in Table 7 are stemmed). The other CS conferences follow expected

trends, with words referring to topics in NLP, Computer Vision, and Machine Learning covering their respective conference. Taken together, these results suggest that AIE conferences are covering the topics that they plan to cover and are highly distinguishable from other conferences.

Conference	Top tokens (w/o stop words, stemmed)
AIE	fair, ai, norm, social, ethic, explan, bia, decis, machin learn, machin
AAAI	plan, intern, fund, european, research, learn, ai, agent, grant, label
ACL	languag, sentenc, semant, corpu, pars, translat, text, grammar, linguist, tag
CVPR	imag, camera, track, line, video, scene, shape, segment, descriptor, face
ECCV	imag, video, scene, object, depth, motion, track, textur, reconstruct, shape
EMNLP	word, languag, sentiment, translat, improv, social media, spanish, media, project, model
HLT-NAACL	word, translat, dialogu, languag, present, parser, spoken, text, dialog, speech
ICCV	imag, object, motion, pose, local, scene, segment, match, surfac, camera
ICML	algorithm, meet, classif, kernel, signal, problem, learn, base, optim, network
NAACL	task, word, subtask, languag, particip, similar, neural, tweet, sentiment, semev
NeurIPS	network, function, learn, distribut, neuron, estim, input, neural, weight, kernel

Table 7. Top words indicative of each conference. AIE: AI Ethics conferences (articles from 2018-present). Stop word and stemming applied before analysis.

5 Discussion

In this work, we have investigated the characteristics and trends of AI Ethics conferences. We have compared them to other CS conferences, such as CVPR, NeurIPS, and ACL. We have found significant differences between AIE and CS conferences in terms of the countries, institutions, and fields involved. We also analyzed differences and temporal dynamics of co-authorship gender composition, field, and country diversities as well as gender, race, and seniority. We also examined the citation between AIE and CS conferences and their possible temporal evolution, showing that AIE conferences seem to start becoming more independent and have an increased impact on other CS conferences. Finally, we examined whether AIE conferences are truly different from other CS conferences by studying how well we can classify papers only based on their content. We show that we can quickly tell apart AIE articles.

While observing the country composition of affiliations for those published in either CS or AIE conferences, we found significant changes in China-affiliated authors. US affiliated authors dominated both kinds of conferences, but Chinese authors do not seem prominent in AIE conferences. We suspect that these stemmed from different priorities and cultural differences as it is known that AI ethics research is not equally distributed worldwide. (Jobin et al., 2019; Roberts et al., 2020). However, we expect that Chinese-affiliated authors will start contributing to AIE conferences more robustly in the future.

We have attempted to understand how AIE and CS conferences differentiate themselves. We have found many areas where they are genuinely different, especially around the language they use (e.g., predictability based on content) and their focus (e.g., field diversity). However, in many ways, both sets of conferences are hard to compare. CS conferences are much older than FAccT and ACM AIES. However, new conferences need to attract the attention of existing scientists.

Moreover, the short existence of the AIE conferences we analyze here can tell a small part of the story. While we can measure changes in non-male and non-white scientists' participation, it is hard to extrapolate. Also, current analytical tools make it hard to differentiate among non-binary genders and other identities. However, in many ways, AIE conferences' current status seems to be at a healthy stage—granted, based on a limited set of analysis suggests.

There are some limitations to how we can take the lessons learned from our analyses into other AIE conferences. If anything, our analysis is displaying what has already been done. It does not show blindspots in what the AIE needs. For example, our analysis of fields present in AIE (Table 5) is based on fields that already exist in the database and probably existed for many years, if not decades. But it is unclear whether and when we need new fields. The tokens that are most predictive of AIE conferences (Table 7) show that fairness, social, and ethical aspects of artificial intelligence should perhaps be part of an entirely new field. Previous research has shown that when new fields emerge in science, they are based on theoretical foundations that generalize a set of phenomena [47]. Perhaps AIE should strive to provide these theories about the relationship between AI in socio-technical systems. Future work should possibly explore the extent to which these new theories are already emerging in this area.

Our analysis here can guide us to understand the emergent field of ethics in artificial intelligence because our results point to a significant difference between AIE and CS conferences. We can use these results to focus on expanding the topics that we can cover in these conferences. For example, we can incorporate more prominently the topics not present in Table 7 analysis. In particular, the topics of Accountability and Transparency are not at the top of the list. This fact could inform the efforts to focus on the future versions of AIE conferences. Similarly, our results suggest that we could expand the conferences' topics or perhaps even bring awareness of the topic to other fields such as robotics, finance, and law. In sum, our analysis shows promising avenues of further development and sharpening of these conferences.

There are some motivations behind the reasoning of this study that might need further exploration. First, we draw ideas from political theory and liberalism. Broadly speaking, a system is just—in the terms discussed by the philosopher John Rawls [17], [18]—if those affected by the system agree to be subjected. Rawls calls this a *well-ordered society*, which requires *all* its members to understand the principles of justice. This definition of justice is in contrast to the concept of *utilitarianism* in which a system could be just if it achieves maximal benefit—even to the detriment of individual members of society. Taking these somewhat abstract ideas into our study, we should study how these AI systems are fair by making all society members participate in their implementation. More concretely, AI ethics conferences should strive to invite members of underrepresented groups or genders or both. However, there is a flaw in our analysis: we are merely analyzing *researchers* attending a conference. Researchers are generally the elites of societies and do not necessarily represent citizens who receive unfair AI systems treatment. After all, researchers are at the high-end distribution of income and many other factors [48], [49]. An alternative solution could be to actively invite ordinary members (i.e., non-researchers) of groups disproportionately affected by AI biases to these conferences. How would someone wrongly convicted by an AI system enrich our understanding of injustices and the consequences of our designed systems? There is a rich literature on reparations that favors such an approach [50]. Hence, these preliminary ideas indicate that much further and broader work needs to be explored in the future.

6 Conclusion

In this article, we have explored how AIE and CS conferences differ. We have analyzed them under a broad set of features related to authors, institutions, fields, countries, and citations. We found that indeed AIE conferences are significantly different from other CS conferences and that they keep features that perhaps are undesirable. For example, they could improve country, race, and gender diversity.

As more versions of the conferences emerge, we will have a better understanding of how AI ethics is evolving. Our research could serve as a guide to explore new topics and serve communities of researchers working in AI ethics but are still not served by CS conferences. The hope is that the methodology, datasets, and questions explored in our research can help as a springboard for guidance.

Acknowledgements

The authors were partially funded by NSF grant #1933803 “Social Dynamics of Knowledge Transfer Through Scientific Mentorship and Publication”. The authors would like to thank the anonymous reviewers for their insightful comments.

References

- [1] Yaser S. Abu-Mostafa, “The first NIPS/NeurIPS,” 2021. <http://work.caltech.edu/neurips.html> (accessed Jan. 29, 2021).
- [2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd edition. Upper Saddle River: Pearson, 2009.
- [3] C. O’Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, First edition. New York: Crown,

- 2016.
- [4] A. Jobin, M. Lenca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, Sep. 2019, doi: <https://doi.org/10.1038/s42256-019-0088-2>.
 - [5] D. Schiff, J. Biddle, J. Borenstein, and K. Laas, "What's Next for AI Ethics, Policy, and Governance? A Global Overview," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Feb. 2020, pp. 153–158, doi: 10.1145/3375627.3375804.
 - [6] D. Serwadda, P. Ndebele, M. K. Grabowski, F. Bajunirwe, and R. K. Wanyenze, "Open data sharing and the Global South—Who benefits?," *Science*, vol. 359, no. 6376, pp. 642–643, Feb. 2018, doi: 10.1126/science.aap8395.
 - [7] S. Fortunato *et al.*, "Science of science," *Science*, vol. 359, no. 6379, Mar. 2018, doi: 10.1126/science.aao0185.
 - [8] D. Crevier, *AI: The Tumultuous History Of The Search For Artificial Intelligence*. NY: Basic Books, 1993.
 - [9] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, 2018.
 - [10] J. A. Mattu Jeff Larson, Lauren Kirchner, Surya, "Machine Bias," *ProPublica*, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=LrqwtD3z1Jth8ag9cay6c0yzKoghtu9C> (accessed Oct. 08, 2020).
 - [11] M. Hurley and J. Adebayo, "Credit Scoring in the Era of Big Data," *Yale Journal of Law and Technology*, 2017, Accessed: Oct. 08, 2020. [Online]. Available: <https://yjolt.org/credit-scoring-era-big-data>.
 - [12] R. B. Freeman and J. Furman, "The Great AI/Robot Jobs Scare: Reality or ... Not Reality of Automation Fear Redux," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans LA USA, Dec. 2018, pp. 1–1, doi: 10.1145/3278721.3278805.
 - [13] S. A. Friedler and C. Wilson, "Preface," in *Conference on Fairness, Accountability and Transparency*, Jan. 2018, pp. 1–2, Accessed: Jan. 29, 2021. [Online]. Available: <http://proceedings.mlr.press/v81/friedler18a.html>.
 - [14] Joshua Kroll and Suresh Venkatasubramanian, "ACM FAccT - 2018 Information for Press," 2018. https://facctconference.org/2018/press_release.html (accessed Jan. 29, 2021).
 - [15] L. Liang and D. E. Acuna, "Artificial mental phenomena: psychophysics as a framework to detect perception biases in AI models," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 403–412, doi: 10.1145/3351095.3375623.
 - [16] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Jan. 2019, pp. 289–295, doi: 10.1145/3306618.3314243.
 - [17] J. Rawls, *A Theory of Justice*, 2nd edition. Cambridge, Mass: Belknap Press: An Imprint of Harvard University Press, 1999.
 - [18] J. Rawls, *Political Liberalism*, Expanded edition. New York: Columbia University Press, 2005.
 - [19] R. Muldoon, *Social contract theory for a diverse world: Beyond tolerance*. Taylor & Francis, 2016.
 - [20] J. M. Cohoon, Z. Wu, and J. Chao, "Sexism: toxic to women's persistence in CSE doctoral programs," *SIGCSE Bull.*, vol. 41, no. 1, pp. 158–162, Mar. 2009, doi: 10.1145/1539024.1508924.
 - [21] S. Cheryan, V. C. Plaut, C. Handron, and L. Hudson, "The Stereotypical Computer Scientist: Gendered Media Representations as a Barrier to Inclusion for Women," *Sex Roles*, vol. 69, no. 1, pp. 58–71, Jul. 2013, doi: 10.1007/s11199-013-0296-x.
 - [22] P. Azoulay, C. Fons-Rosen, and J. S. Graff Zivin, "Does Science Advance One Funeral at a Time?," *American Economic Review*, vol. 109, no. 8, pp. 2889–2920, Aug. 2019, doi: 10.1257/aer.20161574.
 - [23] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, pp. 201–2, Sep. 2012, doi: 10.1038/489201a.
 - [24] J. Terrell *et al.*, "Gender differences and bias in open source: pull request acceptance of women versus men," *PeerJ Comput. Sci.*, vol. 3, p. e111, May 2017, doi: 10.7717/peerj-cs.111.
 - [25] E. Horvitz, "One-Hundred Year Study on Artificial Intelligence: Reflections and Framing," 2014. <https://ai100.stanford.edu/reflections-and-framing> (accessed Oct. 08, 2020).
 - [26] S. Lowry and G. Macpherson, "A blot on the profession," *Br Med J (Clin Res Ed)*, vol. 296, no. 6623, pp. 657–658, Mar. 1988.
 - [27] J. C. Giarratano and G. Riley, *Expert Systems*, 3rd ed. USA: PWS Publishing Co., 1998.
 - [28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.
 - [29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Machine Learning series)*. 2015.
 - [30] M. Herrera, D. C. Roberts, and N. Gulbahce, "Mapping the Evolution of Scientific Fields," *PLOS ONE*, vol. 5, no. 5, p. e10355, May 2010, doi: 10.1371/journal.pone.0010355.
 - [31] M. Krenn and A. Zeilinger, "Predicting research trends with semantic and neural networks with an application in quantum physics," *PNAS*, vol. 117, no. 4, pp. 1910–1916, Jan. 2020, doi: 10.1073/pnas.1914370116.
 - [32] A. Zeng *et al.*, "The science of science: from the perspective of complex systems," *undefined*, 2017. .
 - [33] A. Sinha *et al.*, "An Overview of Microsoft Academic Service (MAS) and Applications," in *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, Florence, Italy, 2015, pp. 243–246, doi: 10.1145/2740908.2742839.
 - [34] S. Fricke, "Semantic Scholar," *jmla*, vol. 106, no. 1, Jan. 2018, doi: 10.5195/JMLA.2018.280.
 - [35] T. Vette, "Genni + Ethnea for the Author-ity 2009 dataset." 2018, [Online]. Available: https://doi.org/10.13012/B2IDB-9087546_V1.
 - [36] Social Security Administration, "Popular Baby Names," 2013. <https://www.ssa.gov/oact/babynames/limits.html> (accessed May 05, 2021).
 - [37] A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena, "Name-ethnicity classification from open sources," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2009, pp. 49–58.
 - [38] Demografix ApS, "Genderize.io | Determine the gender of a name," 2021. <https://genderize.io/> (accessed May 05, 2021).
 - [39] I. S. Pérez, *gender-guesser: Get the gender from first name*. 2016.
 - [40] Gender-API.com, "Gender API," *Gender API*, 2014. <https://gender-api.com/> (accessed May 05, 2021).
 - [41] L. Santamaría and H. Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Comput. Sci.*, vol. 4, p. e156, Jul. 2018, doi: 10.7717/peerj-cs.156.
 - [42] M. Schreiber, "An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the A-index, and the R-index," *J. Am. Soc. Inf. Sci.*, vol. 59, no. 9, pp. 1513–1522, Jul. 2008, doi: 10.1002/asi.20856.
 - [43] A. Thampi, *reverse-geocoder*. 2016.
 - [44] M. F. Porter, "An algorithm for suffix stripping," *Program*, no. 3, pp. 130–137, 1980.
 - [45] S. Cunningham, *Causal Inference: The Mixtape*. New Haven: Yale University Press, 2021.
 - [46] H. Roberts, J. Cows, J. Morley, M. Taddeo, V. Wang, and L. Floridi, "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation," *AI & Society*, Jun. 2020, doi: <https://doi.org.libezproxy2.syr.edu/10.1007/s00146-020-00992-2>.
 - [47] L. Darden, "Discoveries and the Emergence of New Fields in Science," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 1978, no. 1, pp. 149–160, Jan. 1978, doi: 10.1086/psaprocbienmeetp.1978.1.192633.
 - [48] J. R. Behrman and N. Stacey, *The Social Benefits of Education*. University of Michigan Press, 1997.
 - [49] E. Torpey, "Measuring the value of education: Career Outlook: U.S. Bureau of Labor Statistics," 2018. <https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm> (accessed Jan. 31, 2021).
 - [50] A. Buti, "The Notion of Reparations as a Restorative Justice Measure," in *One Country, Two Systems, Three Legal Orders - Perspectives of Evolution*, Berlin, Heidelberg, 2009, pp. 191–206, doi: 10.1007/978-3-540-68572-2_10.