Report on the CyCAT Winter School on Fairness, Accountability, Transparency and Ethics (FATE) in AI

Styliani Kleanthous
CyCAT
Open University of Cyprus
styliani.kleanthous@ouc.ac.cy

Jahna Otterbacher
CyCAT
Open University of Cyprus
jahna.otterbacher@ouc.ac.cy

Jo Bates
University of Sheffield
jo.bates@sheffield.ac.uk

Fausto Giunchiglia
University of Trento
fausto@disi.unitn.it

Frank Hopfgartner
University of Sheffield
f.hopfgartner@sheffield.ac.uk

Tsvi Kuflik
University of Haifa
tsvikak@is.haifa.ac.il

Kalia Orphanou
CyCAT, Open University of Cyprus
kalia.orphanou@ouc.ac.cy

Monica L. Paramita
University of Sheffield
m.paramita@sheffield.ac.uk

Michael Rovatsos
University of Edinburgh
Michael.Rovatsos@ed.ac.uk

Avital Shulner-Tal
University of Haifa
avitalshulner@gmail.com

Abstract

The first FATE Winter School, organized by the Cyprus Center for Algorithmic Transparency (CyCAT) provided a forum for both students as well as senior researchers to examine the complex topic of Fairness, Accountability, Transparency and Ethics (FATE). Through a program that included two invited keynotes, as well as sessions led by CyCAT partners across Europe and Israel, participants were exposed to a range of approaches on FATE, in a holistic manner. During the Winter School, the team also organized a hands-on activity to evaluate a tool-based intervention where participants interacted with eight prototypes of bias-aware search engines. Finally, participants were invited to join one of four collaborative projects coordinated by CyCAT, thus furthering common understanding and interdisciplinary collaboration on this emerging topic.

1 Introduction

Algorithms and analytics play an increasing role in information access, underlying popular information services and social media. Algorithms allow the exploitation of rich and varied data sources, to support human decision-making and/or take direct actions; however, there are increasing concerns surrounding their "social behaviors." There is growing recognition that even when designers and engineers have the best of intentions, systems relying on algorithmic processes can inadvertently result in serious consequences in the social world, such as biases in their outputs that can result in discrimination against individuals and/or groups of people.

With this in mind, it is important to raise awareness of Fairness, Accountability, Transparency and Ethics (FATE) issues in a higher education context. There have been increasing calls for these aspects to be included in the curriculum and training of data scientists in general [Danyluk et al., 2021]. However, as studies (e.g., [Bates et al., 2020]) have shown, this comes with its own challenges.

The EU-H2020 funded Twinning project CyCAT (Cyprus Center for Algorithmic Transparency, Grant Agreement No. 810105) aims to contribute to the education aspect by organising a winter school for educating students and researchers on the issues of FATE in AI systems. The winter school [1] was titled Fairness, Accountability, Transparency and Ethics (FATE) in AI, and is one of the planned activities of the CyCAT project.

2 Goals and Scope of the Winter School

The School targeted Postgraduate students at Masters and PhD levels, and early career postdoctoral researchers. Topics covered include lectures related to algorithmic systems, with the aim to promote the multidisciplinarity of FATE, in the spirit of human-centered Artificial Intelligence (AI). The goal of the FATE in AI Winter School was to bring people into a holistic environment, with participants, as well as speakers, coming from different disciplines, allowing them to get a different perspective than the one they hold, which would result in new insights on the topics of FATE. Our objective was to educate participants, broaden their knowledge and emphasize the importance of FATE topics by gaining the experience and creating resources for providing awareness. The knowledge participants' gained was demonstrated through the hands-on activities they worked on after the winter school.

A total of 136 participants pre-registered for the Winter School (48 female, 87 male). Their geographical distribution (in terms of their country of work) is shown in Figure 1. As can be seen, most participants were located in Europe or Israel, while 17% were located in the U.S. The majority of the participants indicated in their registration a university affiliation, although a small number (n=17) listed an industry or government affiliation.

[1] https://sites.google.com/view/cycat-winter-school/home

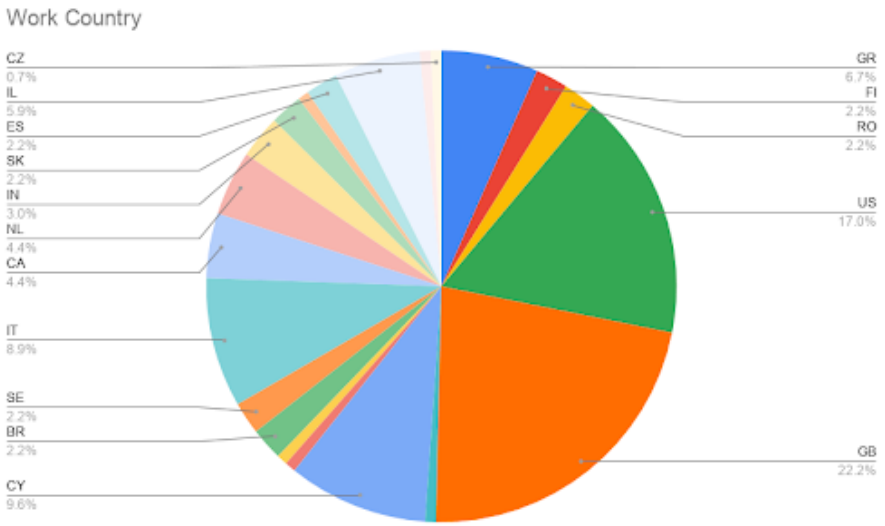

Figure 1: Countries in which registered participants work.

3 Sessions

3.1 Keynotes

- **Joanna J. Bryson, Hertie School of Governance (Berlin, Germany)**
  Personal and Transnational Economic Impacts of AI and Digital Technology
  Digital technologies and particularly artificial intelligence change the landscape through which humans express our identity – our security, our economy, and our culture. In this lecture particular attention was paid to wages and the future of work, and to inequality and its concomitant polarisation, showing recent research outcomes in each. These outcomes were reflected on from a deeply human-centred perspective, discussing not only potential positive and negative outcomes, but the limits of how much can change.

- **Casey Dugan, IBM Research, Cambridge Research Center (Massachusetts, USA)**
  Human-Centered AI at IBM Research - Automation versus Collaboration in the Age of AI
  Advances in artificial intelligence create unprecedented opportunities for automating tasks that previously could only be done by humans - for example, driving cars, writing essays, or creating novel drugs – but also impacts the workers doing those tasks today (i.e. rideshare drivers, authors, and chemists). As AI is infused into more and more systems, there is increasing need to study and understand the impact automation has on our workforce, but also on how we interact with intelligent systems and how such systems need to be designed to most effectively support their human users. Our Human-Centered AI agenda at IBM Research has explored a different perspective on Human-AI interaction, investigating the transformation of the interaction to a more collaborative relationship in which humans and AI systems work hand-in-hand to create a desired outcome. In this talk the theme of collaboration versus automation with AI was explored through a number of research projects

and scientific studies that were conducted in the context of AI Lifecycle Management, Automated Model Generation and Exploration for Data Scientists, Human-in-Loop Data Labeling, Explainability and Trust in AI systems, and AI-Infused Process Automation, as well as Generative Models and how they will fundamentally change how humans will interact with AI systems in the creative industries and for content generation. This talk gave a unique industry perspective on designing and building AI systems with users in mind.

3.2 FATE Topics

- **Prof. Michael Rovatsos, University of Edinburgh**
  AI Ethics
  As the impact of AI on society is increasing rapidly, a breadth of ethical concerns are surfacing in the current debate on AI, and we are witnessing a transition from speculative explorations of its potential future impact on humanity to concrete ethical challenges encountered by businesses and governments as they are adopting AI technologies in real-world systems. This talk provided an overview of the current landscape in AI ethics covering both philosophical and societal questions as well as some examples, practical methodologies and technical approaches that have been proposed to tackle some of these emerging problems.

- **Prof. Fausto Giunchiglia, University of Trento**
  Diversity, bias and related issues
  The virtualization of the world has generated huge amounts of data which, in turn, have fueled a lot of data driven applications, many of which are influencing the life of people, e.g., user profiling, search/ query answering, task and job allocation. While generating a lot of positive effects, this phenomenon has soon also shown some inherent limitations, one of which is that the systems show bias, with respect to, e.g., gender, race, opinions and many more dimensions. Various partial solutions to this problem have been provided, e.g., (algorithmic) transparency, explainability, fairness. In this talk Prof. Giunchiglia argued that the problem of bias is a consequence of the fact that data generators, application developers, and application users all live in different contexts and that, as such, bring diverse perspectives. As outlined in the paper by Giunchiglia et al. [2021], this diversity in the perspective is the unavoidable source of bias. The solution proposed is that of making data provenance explicit as the necessary condition for providing generality to the solutions provided so far. Here by "data provenance" we mean keeping trace of all the contexts within which data are generated, manipulated and used.

- **Prof. Tsvi Kuflik, University of Haifa**
  End-Users' Perception of Algorithmic Fairness
  Machine Learning (ML) and Artificial Intelligence (AI) based systems ("Algorithmic Systems") are starting to become part of our everyday lives. They are used to filter job applications, credit requests, judicial and medical decisions, driving safety and many more controversial applications. These systems are usually considered to be "black-box" systems, that their reasoning and results are not always clear to their users and that they are not always fair to their users. The lack of explanations about how a system works and how and why decisions/recommendations were made may bring with them the risk of undetected

bias, discrimination and perception of unfairness. As a result, the issue of transparency and fairness of such algorithmic systems draws a lot of research attention in recent years, as the harmful potential of biased algorithms has been recognized by researchers and practitioners.

Accordingly, we have also witnessed a growing interest in ensuring the fairness and transparency of such systems. Yet, so far, there is no agreed upon solution and not even an agreed upon terminology. Recent research focuses mainly on formal verification of fairness and there is a lack of studies about perceived fairness and its impact on users' decisions to use a system and to trust its results. Hence, we need to understand stakeholders' fairness perception regarding algorithmic systems. This can be done by exploring what impacts users' fairness perception and finding ways to measure the perceived fairness. This, in turn can become part of the development of fair and transparent algorithmic systems.

The talk discussed the challenges posed by these "black boxes" and state of the art ideas for solutions all in a framework of a holistic model of algorithmic fairness.

- **Ms. Casey Dugan, IBM**
  The intersection of AI and HCI: Gamifying the latest artificial intelligence research
  AI is increasingly being incorporated into critical applications impacting our everyday lives and society at large, from self-driving cars to parole decisions. However, machine learning classifier is susceptible to a backdoor attack, or free from bias. In this talk, Ms. Dugan described the work done by IBM Research in developing innovations that help create AI technologies that are trustworthy, i.e. unbiased, robust, explainable, and transparent. The advances in these areas have been shared as open source, as well as published papers, and even games. Casey Dugan's team has created Learn and Play, a series of web-based games that educate the public about new AI technologies.

  Ms. Dugan further described the Design Thinking process to build these games within small agile teams, how they jump start development with both IBM and external open source, and deploy and test with end users. Participants learned how they can leverage these open source tools and resources to make their own AI-infused applications more trustworthy and contribute to the ecosystem for advancing these crucial technologies.

- **Dr. Frank Hopfgartner and Dr. Jo Bates, University of Sheffield**
  Bias and Transparency of Web Search Engines
  Web search engines play an important role in our daily information gathering routine. With the global market mostly dominated by a few Web search engine providers this leaves a lot of power in the hands of very few. Considering this situation, it is important to make people aware of potential biases and emphasize the need for transparency in Web search engines. In this lecture, the speakers approached these challenges from two perspectives. Participants were first introduced to basic technical concepts underlying Web search such as document crawling, indexing, and ranking. Further, a brief introduction to common personalization approaches such as [Vallet et al., 2008] was given. This was then followed by a critical reflection on societal challenges arising from personalized content filtering techniques and an illustration of the lack of transparency. Several case studies were presented that highlight the role of search engines when dealing with biased content.

- **Dr. Jahna Otterbacher, CyCAT, Open University of Cyprus**
  Bias in Data and Algorithmic Systems: Problems, Solutions and Stakeholders
  Mitigating bias in algorithmic processes and systems is a critical issue drawing increasing attention across research communities within the information and computer sciences. Given the complexity of the problem and the involvement of multiple stakeholders – not only developers, but also end-users and third parties – there is a need to understand the landscape of the sources of bias, as well as the solutions being proposed to address them. In this talk, Dr. Otterbacher presented insights from a recent survey of 250+ articles across four domains (machine learning, information retrieval, HCI, and RecSys), providing a "fish-eye view" of the field [Orphanou et al., 2021]. In the second part of the talk examples of previous work done by the CyCAT project on auditing proprietary computer vision systems for social biases, positioning this work vis-à-vis the aforementioned framework as well as the emerging science of machine behavior were discussed.

- **Dr. Styliani Kleanthous, CyCAT, Open University of Cyprus**
  Perceptions of Young Developers on Algorithmic Fairness, Transparency and Accountability
  Algorithmic decision-making systems are becoming very popular, prompting us to rely more and more on their decisions, with potentially serious consequences for the affected social groups. Developers have an important role to play when they are called to develop algorithms that will drive these decisions. Algorithmic fairness might be a first step in understanding how people perceive and assess the decisions and the explanations provided. Most importantly, we need to understand how developers perceive fairness in the systems they develop, which will potentially decide on behalf of a human, and in some occasions for matters with real social impact.

  This talk provided some insights on how future developers perceive algorithmic fairness in algorithmic decision-making [Kasinidou et al., 2021]. It looked into the role that academic education has to play in their understanding of the decision-making process, as well as their critical thinking on the factors and the decision-making process involved.

3.3 Collaborative Projects

Here, we provide brief summaries of the collaborative projects that were designed and led by the CyCAT Consortium members. Students could optionally join one of the projects, which were launched on the final day of the School. The teams worked together over the course of six weeks, and presented their results in an online seminar, held at the end of February 2021.

Diversity or Bias? Using transparency paths for self-reflection. The aim of this project was to put in practice the ideas presented in Fausto's Giunchiglia talk, *Diversity, Bias and Related Issues*. The key idea was to study how much the process by which bias is perceived is influenced by the underlying diversity in the world and in the mental or Web representations that people build about the world. The final goal was to identify *transparency paths*, namely the steps by which the diversity and, possibly the sources of bias, were generated.

State-of-the-Art in Explainable Artificial Intelligence (XAI). The aim of this project was to review recent literature about explanations of algorithmic systems starting from 2016, the

year in which the right to explanation of algorithmic systems was presented (within GDPR), to analyze and identify interesting aspects of explanations and to create a summary of the various explanation styles that are used in such systems. Six categories were identified, and used to classify the existing work in the area.

Comparing Tools for Bias Testing and Algorithm Auditing. Being able to identify potential biases in data and algorithms has become an important area of focus for researchers and practitioners. There have been several tools and libraries created for bias testing and algorithm auditing, many of which are open source and publicly available. Despite their importance, no critical analysis of their scope or functionalities have been performed yet. The main aim of this project was to provide an overview of open source tools and libraries that have been developed for bias testing. The participants performed a thorough literature review to identify relevant tools, identified a list of popular open datasets used for research, and identified a methodology to evaluate and compare these tools on the identified datasets.

Designing a methodology to monitor bias in recruitment platforms. The project covered the topic of monitoring bias in algorithmic tools used in recruitment. We selected LinkedIn's job search function as a specific use case. The participants identified the types of data that job search results should be based on, such as relevant data including skills, education and experience. Then the basic functions for recruitment platforms were analyzed. The project resulted in a set of high-level recommendations and mitigation proposals.

3.4 Hands-on Activity: Tool-Based Intervention

In addition to the collaborative projects, we also organized a 2-hour session to discuss how bias should be addressed in search engines. This event was attended by 16 participants, ranging from postgraduate students, to academics and practitioners from ten countries worldwide (Brazil, US, India, and seven countries in Europe). In this session, participants were asked to interact with eight prototypes of bias-aware systems, each provided users with different features and level of controls. Participants worked in small groups and were asked to discuss how they would rank the systems based on their preferences and to reflect on how these approaches would influence their information seeking tasks. They were then invited to discuss their findings with the whole group. This session raised an interesting discussion across all participants on how search engines should address biases in search engines.

4 Discussion and Conclusion

The COVID-19 pandemic challenged us to design and deliver an engaging virtual event. During the event in early January 2021, most participants – as well as consortium members – indicated that they were in a lockdown situation. Despite the challenges, our Winter School had a successful turnout, with over 100 participants registered and approximately 60 online within any given session. Providing a mix of cutting-edge talks, hands-on activities, as well as social breaks, helped to keep the event lively and well-attended.

We believe that the FATE Winter School aided in increasing young researchers' awareness of the challenges of algorithmic bias, at the same time introducing them to tools, frameworks and methodologies. In addition, it provided an opportunity to interact with more senior researchers

and to get involved in collaborative projects, where small groups of researchers and students continued to work independently, focusing on specific topics of interest, deepening their knowledge of the subject. The results of these small group sessions were presented at a concluding session of the winter school. Moreover, the collaborative work continued even after the official end of the winter school and two of which have resulted in papers that were accepted to a related workshop. In a post-event questionnaire, most respondents indicated that they would be very interested in attending similar virtual events and schools in the future.

In response to an open-ended question, "What was the best thing about the Winter School for you?" participants had varied responses, such as the following:

- "The panel discussions!"
- "Learn more about FATE in AI, to know what type of research are being conducted, the warm welcoming and open space for making questions and comments, and the opportunity to work with a research group outside my regular circle."
- "The whole programme. Most talks left me with many thoughts regarding my own PhD project, talks and panel discussions were very interesting and inspiring and motivated me a lot."
- "That it was online and free making it accessible to everyone. The talks are all very interesting and important."

In conclusion, this was the first iteration of our FATE Winter School, and as we enrich our methodologies for educating young researchers, and as it becomes possible to conduct events in a face-to-face and/or hybrid format, the approach to FATE education in this format will evolve. We would encourage readers (i.e., our fellow academics) to take FATE to heart, incorporating similar approaches, in their efforts to education the next generation of researchers and practitioners.

Acknowledgements

This research has been supported by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105 (CyCAT: Cyprus Center for Algorithmic Transparency).

References

Jo Bates, David Cameron, Alessandro Checco, Paul Clough, Frank Hopfgartner, Suvodeep Mazumdar, Laura Sbaffi, Peter Stordy, and Antonio de la Vega de León. Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 425–435, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372832.

Andrea Danyluk, Paul Leidig, Andrew McGettrick, Lillian Cassel, Maureen Doyle, Christian Servin, Karl Schmitt, and Andreas Stefik. Computing competencies for undergraduate data science programs: an acm task force final report. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1119–1120, 2021.

Fausto Giunchiglia, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Veronika Bogina, Tsvi Kuflik, and Avital Shulner Tal. Towards algorithmic transparency: A diversity perspective. *arXiv*, 2021. URL https://arxiv.org/abs/2104.05658.

Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 690–700, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445931.

Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Tsvi Kuflik, et al. Mitigating bias in algorithmic systems: A fish-eye view of problems and solutions across domains. *arXiv*, 2021. URL https://arxiv.org/abs/2103.16953.

David Vallet, Frank Hopfgartner, and Joemon M. Jose. Use of implicit graph for recommending relevant videos: A simulated evaluation. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 of *Lecture Notes in Computer Science*, pages 199–210. Springer, 2008. doi: 10.1007/978-3-540-78646-7_20.