Report on the 2019 International Joint Conferences on Artificial Intelligence Explainable Artificial Intelligence Workshop

Tim Miller, Rosina Weber, Dan Magazzeni

■ *This article reports on the Explainable Artificial Intelligence Workshop, held within the International Joint Conferences on Artificial Intelligence 2019 Workshop Program in Macau, August 11, 2019. With over 160 registered attendees, the workshop was the largest workshop at the conference. It featured an invited talk and 23 oral presentations, and closed with an audience discussion about where explainable artificial intelligence research stands.*

As artificial intelligence (AI) becomes more ubiquitous, complex, and consequential, the need for people to understand how decisions are made and to judge their correctness becomes increasingly crucial due to concerns of ethics, accountability, and trust. The field of explainable AI (XAI) aims to address this problem by designing AI whose decisions can be understood by humans. The workshops in XAI have been receiving growing interest. The 2019 International Joint Conferences on Artificial Intelligence's Explainable Artificial Intelligence workshop attracted 163 registered attendees, following the tradition of being the largest International Joint Conferences on Artificial Intelligence workshop since 2017.

Copyright © 2020, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

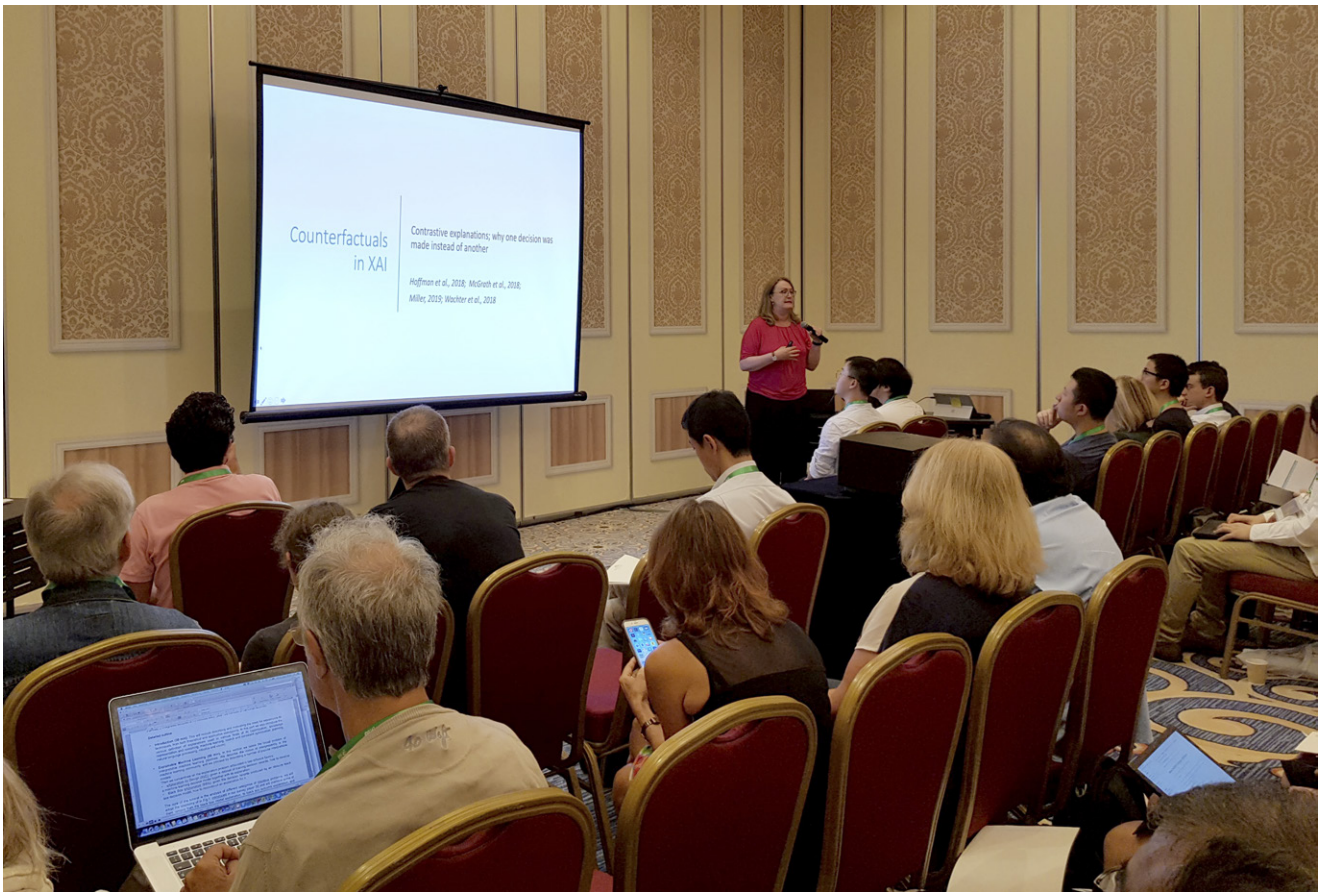

Figure 1. Professor Ruth Byrne Delivering the XAI 19 Workshop Keynote.

We received fifty-five submissions in total. Each submission received two reviews, and one metareview assessing quality and potential interest. In the end, twenty-three papers were accepted into the workshop. Different from previous years where there were poster sessions and multiple invited talks, this year's workshop only had oral presentations. The workshop proceedings and presentations are publicly available.[1]

The workshop began with an invited talk given by Ruth Byrne (Trinity College Dublin) who spoke about counterfactuals (figure 1). Byrne's International Joint Conferences on Artificial Intelligence 2019 reviews discusses the implications for XAI from works in psychology and cognitive science such as how humans create, interpret, and use counterfactuals.[2] The hour-long invited talk allowed us time to have a detailed presentation of her survey compared with her main track presentation. This talk covered many interesting and useful bodies of work with which anyone working in XAI should be familiar.

The contents of this year's workshop can be roughly clustered in three major topics, namely reinforcement learning (RL), XAI and humans, and others. The interest in RL is significant. RL was the most frequent theme in this year's workshop presentations — unsurprising given the recent interest in deep RL. Explainable RL presents challenges not seen in most supervised and unsupervised problems because RL models a sequential decision-making process. However, many of the techniques presented build on earlier work in supervised learning, such as distilling policies into decision trees, and using saliency maps to highlight important features.

The second group, despite targeting different AI methods, either evaluated explanations with human studies or focused on collaboration between humans and autonomous agents. The fact that explainability is typically defined in terms of impact on humans[3] motivates attention to human factors and human-centered evaluations. The works in this group includes explanations for hints on intelligent tutoring systems,

where authors ask students which explanations they want. One study showed certain abstractions to explain unsolvability match human intuition, while another investigated the impact of prediction errors, and one investigated the long-standing issue in XAI of persuasive but potentially incorrect explanations. The latter found that fidelity is crucial to user trust. Human-machine collaboration was discussed from the perspective of humans giving feedback to improve machine learning accuracy and to increasing trust when making decisions under uncertainty.

The final group of presentations consisted of works in diverse areas such as generic frameworks for XAI and for gathering requirements for XAI, transparency obtained from rules and case-based reasoning, and temporal considerations. There were also papers focusing on images and visualizations for XAI.

The workshop concluded with a discussion among all attendees. During the discussion, a few observations emerged. It was noted that evaluation of XAI methods has come a long way in just two years. In the 2017 XAI workshop, for example, evaluation was scarce, and there were no reports of human subject evaluations. By 2019, however, most papers had at least some computational evaluation, and many had human subject evaluations or were already working on them.

Nevertheless, it was also noted that evaluation remains a challenge. One observation was that that subjective nature of concepts such as explainability and interpretability may motivate authors to propose their own specific working definition for evaluation. The result is that evaluations can be ad hoc, making it very difficult to assess generalized progress in the field. The conclusion was that more work is required.

For human-factors studies, authors use different numbers of subjects and perform quite different studies, such as surveys versus interviews, and studies where users are asked questions about events versus others where they are asked to execute some task. What could help here are clear examples around types of evaluations for particular types of problems, although it remains to be seen whether this is possible.

Overall, the workshop was a resounding success, with many quality submissions and presentations. Our current desire is to continue this workshop with another iteration at International Joint Conferences on Artificial Intelligence 2020 in Japan.[4]

Notes

1. sites.google.com/view/xai2019/home
2. www.ijcai.org/proceedings/2019/0876.pdf
3. See T. Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267: 1–38. February 2019.
4. www.ijcai20.org/

First Call for Nominations for the AAAI Squirrel AI Award

The AAAI Squirrel AI Award for Artificial Intelligence for the Benefit of Humanity recognizes positive impacts of artificial intelligence to protect, enhance, and improve human life in meaningful ways with long-lived effects. The award will be given annually at the conference for the Association for the Advancement of Artificial Intelligence (AAAI, and is accompanied by a prize of $1,000,000 plus travel expenses to the conference. Financial support for the award is provided by Squirrel AI.

Candidates may be individuals, groups, or organizations that are directly connected with the main contribution stated in the nomination. Qualifications or technical knowledge in artificial intelligence are not requirements for nominations. The emphasis is on the significance and impact of the work.

The inaugural award will be given at AAAI-21 in Vancouver, Canada in February 2021. Nominations are due May 24, 2020, End of Day, Anywhere on Earth (AoE), UTC -12 hrs. Nomination packets should be emailed to the AAAI Executive Director at aaai-exec-director @aaai.org. For complete information about nomination requirements, please see www.aaai.org/ Awards/squirrel-award-call.php.

Tim Miller is an associate professor in Computer Science in the School of Computing and Information Systems at The University of Melbourne, where he leads the AI and Autonomy Lab. His primary area of expertise is in AI, with particular emphasis on human-AI interaction and XAI. His work is at the intersection of AI, interaction design, and cognitive science and psychology.

Rosina Weber is an associate professor in the College of Computing and Informatics, Information Science, at Drexel University. Her research interests lie in processing unstructured data (text) to obtain representations to equip AI systems to make decisions based on such data, with particular emphasis on using case-based reasoning and XAI.

Daniele Magazenni is a reader (associate professor) in AI at the Department of Informatics, King's College London, where he leads the Human-AI Teaming laboratory, and is a codirector of the United Kingdom Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence. His research interests are in safe, trusted, and explainable AI, with a particular focus on AI planning for robotics and autonomous systems, machine and RL, and human-AI teaming.