TExSS: Transparency and Explanations in Smart Systems

Alison Smith-Renner
Decisive Analytics Corporation
USA
alisonmarierenner@gmail.com

Jonathan Dodge
Oregon State University
USA
dodgej@oregonstate.edu

Casey Dugan
IBM Research
USA
cadugan@us.ibm.com

Styliani Kleanthous
Cyrpus Centre for Algorithmic
Transparency, Open University of
Cyprus
Cyprus
styliani.kleanthous@gmail.com

Tsvi Kuflik
Information Systems, The University
of Haifa
Israel
tsvikak@is.haifa.ac.il

Min Kyung Lee
University of Texas at Austin
USA
minkyung.lee@austin.utexas.edu

Brian Y Lim
Department of Computer Science,
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg

Advait Sarkar
Microsoft Research
UK
advait@microsoft.com

Avital Shulner-Tal
Information Systems, The University
of Haifa
Israel
avitalshulner@gmail.com

Simone Stumpf
Centre for HCI Design, City,
University of London
UK
simone.stumpf.1@city.ac.uk

ABSTRACT

Smart systems that apply complex reasoning to make decisions and plan behavior, such as decision support systems and personalized recommendations, are difficult for users to understand. Algorithms allow the exploitation of rich and varied data sources, in order to support human decision-making and/or taking direct actions; however, there are increasing concerns surrounding their transparency and accountability, as these processes are typically opaque to the user. Transparency and accountability have attracted increasing interest to provide more effective system training, better reliability and improved usability. This workshop provides a venue for exploring issues that arise in designing, developing and evaluating intelligent user interfaces that provide system transparency or explanations of their behavior. In addition, we focus on approaches to mitigate algorithmic biases that can be applied by researchers, even without access to a given system's inter-workings, such as awareness, data provenance, and validation.

CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools*;
• **Computing methodologies** → *Machine learning*; *Artificial intelligence.*

KEYWORDS

explanations, visualizations, machine learning, intelligent systems, intelligibility, transparency, fairness, accountability

1 INTRODUCTION

Smart systems that apply complex reasoning for decision making, such as clinical decision support systems, face detection, recommender systems, and autonomous vehicles, are difficult for users to understand, interact with, and trust appropriately [2]. Fairness, accountability and transparency are crucial aspects of intelligent systems, especially as these systems become more prevalent in our day-to-day lives. Textual explanations and graphical visualizations are often provided by systems to give insight into what they are doing and why [3–6]; if models are right for the right reasons, users can be more confident that they will generalize or are operating without bias [1]. However, there are still numerous issues

and problems regarding explanations and algorithm transparency that demand further attention, such as how can we build (good) algorithmic systems, particularly those that demonstrate that they are fair, accountable, and unbiased? What should be included in explanations and how (and to whom) should they be presented? How can we evaluate explanations and their ability to accurately explain underlying algorithms and overall systems' behavior, especially for the goals of fairness and accountability?

2 WORKSHOP OVERVIEW

The TExSS 2021 workshop brings together academia and industry researches to exchange perspectives, approaches, and results (https://explainablesystems.comp.nus.edu.sg/2021/). This workshop is the 4th iteration in the "Explainable Smart Systems" series at IUI (after ExSS18, ATEC19, ExSS19, and ExSS-ATEC20). This workshop includes a keynote, paper panels, and group activities, relating to the theme of transparency and explanations for fairness and social justice. The 19 accepted papers are presented as part of themed paper panel sessions and poster sessions. These papers spanned a wide variety of topics, including:

- Fairness perception
  - J. Schoeffer et al. A study on fairness and trust perceptions in automated decision making.
  - A. Shulner-Tal et al. Fairness, explainability, and what lies in between.
  - J. Tang et al. Understanding how customers attribute accountability in food delivery breakdowns.
- Explanations to non-experts
  - A. Ayobi et al. Machine learning explanations as boundary objects: How AI researchers explain and non-experts perceive machine learning.
  - C. Bove et al. Contextualising local explanations for non-expert users: an XAI pricing interface for insurance.
  - A. Islam and A. Bunt. Investigating explanations that target training data.
  - R. Eardley et al. Explaining complex machine learning platforms to members of the general public.
  - R. Lasarati. AI healthcare system interface: Explanation design for non-expert user trust.
- Justice and related aspects
  - I. Nigman and M. Lee. Interactive and explainable AI for advancing organizational justice in the modern workplace.
  - E. Watkins. The tension between information justice and security: Perceptions of facial recognition and targeting.
- Transparency
  - N. Kwon et al. 3D4ALL: Toward an inclusive pipeline to classify 3D contents.
  - S. Stumpf et al. Design methods for artificial intelligence fairness and transparency.
- Human-AI relationships
  - J. Dodge. Who gets to harness (X)AI? For billion-dollar organizations only.
  - J. Jansen Ferreira and M. de Souza Monteiro. The human-AI relationship in decision making: AI explanations to support people in justifying their decisions.

- Z. Zhang and H. Hußmann. How to manage output uncertainty: Targeting the actual end user problem in interactions with AI.
- Applications
  - O. Alkan et al. Making business partner recommendations more effective: Impacts of combining recommenders and explanations through user feedback.
  - M. Guesmi et al. Open, scrutable and explainable interest models for transparent recommendation.
  - C. Tippet. Smart move? Algorithmic transparency in career transition tools.
  - M. Hossein Jarrahi and M. Haeri. Integrating explainable AI in medical diagnosis through parameterization and implicitization.

3 WORKSHOP ORGANIZERS

The workshop was organized by a committee of ten people, spanning Europe, North America and Asia, who attended the workshop and conference. The organizers include Dr. Alison Smith-Renner (Decisive Analytics Corporation, USA), Dr. Styliani Kleanthous (Open University of Cyprus), Dr. Brian Lim (National University of Singapore), Dr. Advait Sarkar (Microsoft Research, Cambridge, UK), Dr. Simone Stumpf (City, University of London, UK), Jonathan Dodge (Oregon State University, USA), Dr. Tsvika Kuflik (University of Haifa), Avital Shulner-Tal (University of Haifa), Dr. Min Kyung Lee (UT Austin, USA), and Casey Dugan (IBM Research, Cambridge, USA). These workshop organizers have a breadth of experience organizing similar workshops at IUI and other conferences.

REFERENCES

[1] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *International Conference on Intelligent User Interfaces (IUI)*. https://doi.org/10.1145/3301275.3302310
[2] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08*. 227. https://doi.org/10.1145/1378773.1378804
[3] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*. 241–250. https://doi.org/10.1145/358916.358995
[4] Carmen Lacave and Francisco J. Díez. 2002. A review of explanation methods for Bayesian networks. , 107–127 pages. https://doi.org/10.1017/S026988890200019X
[5] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *International conference on Intelligent User Interfaces (IUI)*. 93. https://doi.org/10.1145/1111449.1111475
[6] William Swartout, Cecile Paris, and Johanna Moore. 1991. Explanations in knowledge systems: Design for Explainable Expert Systems. *IEEE Expert* 6, 3 (1991), 58–64. https://doi.org/10.1109/64.87686