

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence

14127

Founding Editor

Jörg Siekmann

Series Editors

Randy Goebel, *University of Alberta, Edmonton, Canada*

Wolfgang Wahlster, *DFKI, Berlin, Germany*

Zhi-Hua Zhou, *Nanjing University, Nanjing, China*

The series Lecture Notes in Artificial Intelligence (LNAI) was established in 1988 as a topical subseries of LNCS devoted to artificial intelligence.

The series publishes state-of-the-art research results at a high level. As with the LNCS mother series, the mission of the series is to serve the international R & D community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings.

Davide Calvaresi · Amro Najjar ·
Andrea Omicini · Reyhan Aydogan ·
Rachele Carli · Giovanni Ciatto · Yazan Mualla ·
Kary Främling
Editors

Explainable and Transparent AI and Multi-Agent Systems

5th International Workshop, EXTRAAMAS 2023
London, UK, May 29, 2023
Revised Selected Papers

Editors

Davide Calvaresi 
University of Applied Sciences and Arts
Western Switzerland
Sierre, Switzerland

Andrea Omicini 
Alma Mater Studiorum, Università di
Bologna
Bologna, Italy

Rachele Carli 
Alma Mater Studiorum, Università di
Bologna
Bologna, Italy

Yazan Mualla 
Université de Technologie de
Belfort-Montbéliard
Belfort Cedex, France

Amro Najjar 
Luxembourg Institute of Science
and Technology
Esch-sur-Alzette, Luxembourg

Reyhan Aydogan 
Ozyegin University
Istanbul, Türkiye

Giovanni Ciatto 
Alma Mater Studiorum, Università di
Bologna
Bologna, Italy

Kary Främling 
Umeå University
Umeå, Sweden

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Artificial Intelligence

ISBN 978-3-031-40877-9

ISBN 978-3-031-40878-6 (eBook)

<https://doi.org/10.1007/978-3-031-40878-6>

LNCS Sublibrary: SL7 – Artificial Intelligence

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

AI research has made several significant breakthroughs that has boosted its adoption in several domains, impacting our lives on a daily basis. Nevertheless, such widespread adoption of AI-based systems has raised concerns about their foreseeability and controllability and led to initiatives to “slow down” AI research. While such debates have mainly taken place in the media, several other research works have emphasized that achieving trustworthy and responsible AI would necessitate making AI more transparent and explainable.

Not only would eXplainable AI (XAI) increase acceptability, avoid failures, and foster trust, but it would also comply with relevant (inter)national regulations and highlight these new technologies’ limitations and potential.

In 2023, the fifth edition of the EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS) continued the successful track initiated in 2019 in Montreal and followed by the 2020 to 2022 editions (virtual due to the COVID-19 pandemic circumstances). Finally, EXTRAAMAS 2023 was held in person and proposed bright presentations, a stimulating keynote (titled “Untrustworthy AI” given by Jeremy Pitt, Imperial College London), and engaging discussions and Q&A sessions.

Overall, EXTRAAMAS 2023 welcomed contributions covering areas including (i) XAI in symbolic and subsymbolic AI, (ii) XAI in negotiation and conflict resolution, (iii) Explainable Robots and Practical Applications, and (iv) (X)AI in Law and Ethics.

EXTRAAMAS 2023 received 26 submissions. Each submission underwent a rigorous single-blind peer-review process (three to five reviews per paper). Eventually, 16 papers were accepted and collected in this volume.

Each paper was presented in person (with the authors’ consent, they are available on the EXTRAAMAS website¹). Finally, The Main Chairs would like to thank the special track chairs, publicity chairs, and Program Committee for their valuable work, as well as the authors, presenters, and participants for their engagement.

June 2023

Davide Calvaresi
Amro Najjar
Andrea Omicini
Kary Främling

¹ <https://extraamas.ehealth.hevs.ch/>.

Organization

General Chairs

Davide Calvaresi	University of Applied Sciences and Arts Western Switzerland, Switzerland
Amro Najjar	Luxembourg Institute of Science and Technology, Luxembourg
Andrea Omicini	Alma Mater Studiorum, Università di Bologna, Italy
Kary Främling	Umeå University, Sweden

Special Track Chairs

Reyhan Aydogan	Ozyegin University, Turkiye
Giovanni Ciatto	Alma Mater Studiorum, Università di Bologna, Italy
Yazan Mualla	UTBM, France
Rachele Carli	Alma Mater Studiorum, Università di Bologna, Italy

Publicity Chairs

Yazan Mualla	UTBM, France
Benoit Alcaraz	University of Luxembourg, Luxembourg
Rachele Carli	Alma Mater Studiorum, Università di Bologna, Italy

Advisory Board

Tim Miller	University of Melbourne, Australia
Michael Schumacher	University of Applied Sciences and Arts Western Switzerland, Switzerland
Virginia Dignum	Umeå University, Sweden
Leon van der Torre	University of Luxembourg, Luxembourg

Program Committee

Andrea Agiollo	University of Bologna, Italy
Remy Chaput	University of Lyon 1, France
Paolo Serrani	Università Politecnica delle Marche, Italy
Federico Sabbatini	University of Bologna, Italy
Kary Främling	Umeå University, Sweden
Davide Calvaresi	University of Applied Sciences and Arts Western Switzerland, Switzerland
Rachele Carli	University of Bologna, Italy
Victor Contreras	University of Applied Sciences and Arts Western Switzerland, Switzerland
Francisco Rodríguez Lera	University of León, Spain
Bartłomiej Kucharzyk	Jagiellonian University, Poland
Timotheus Kampik	University of Umeå, Sweden, Signavio GmbH, Germany
Igor Tchappi	University of Luxembourg, Luxembourg
Eskandar Kouicem	Université Grenoble Alpes, France
Mickaël Bettinelli	Université Savoie Mont Blanc, France
Ryuta Arisaka	Kyoto University, Japan
Alaa Daoud	INSA Rouen-Normandie, France
Avleen Malhi	Bournemouth University, UK
Minal Patil	Umeå University, Sweden
Yazan Mualla	UTBM, France
Takayuki Ito	Kyoto University, Japan
Lora Fanda	University of Geneva, Switzerland
Marina Paolanti	Università Politecnica delle Marche, Italy
Arianna Rossi	University of Luxembourg, Luxembourg
Joris Hulstijn	University of Luxembourg, Luxembourg
Mahjoub Dridi	UTBM, France
Thiago Raulino	Federal University of Santa Catarina, Brasil
Giuseppe Pisano	University of Bologna, Italy
Katsuhide Fujita	Tokyo University of Agriculture and Technology, Japan
Jomi Fred Hubner	Federal University of Santa Catarina, Brazil
Roberta Calegari	University of Bologna, Italy
Hui Zhao	Tongji University, China
Niccolo Marini	University of Applied Sciences Western Switzerland, Switzerland
Salima Lamsiyah	University of Luxembourg, Luxembourg
Stephane Galland	UTBM, France
Giovanni Ciatto	University of Bologna, Italy

Matteo Magnini
Viviana Mascardi
Giovanni Sileno
Sarath Sreedharan

University of Bologna, Italy
University of Genoa, Italy
University of Amsterdam, The Netherlands
Arizona State University, USA

Contents

Explainable Agents and Multi-Agent Systems

Mining and Validating Belief-Based Agent Explanations	3
<i>Ahmad Alelaimat, Aditya Ghose, and Hoa Khanh Dam</i>	
Evaluating a Mechanism for Explaining BDI Agent Behaviour	18
<i>Michael Winikoff and Galina Sidorenko</i>	
A General-Purpose Protocol for Multi-agent Based Explanations	38
<i>Giovanni Ciatto, Matteo Magnini, Berk Buzcu, Reyhan Aydoğan, and Andrea Omicini</i>	
Dialogue Explanations for Rule-Based AI Systems	59
<i>Yifan Xu, Joe Collenette, Louise Dennis, and Clare Dixon</i>	
Estimating Causal Responsibility for Explaining Autonomous Behavior	78
<i>Saaduddin Mahmud, Samer B. Nashed, Claudia V. Goldman, and Shlomo Zilberstein</i>	

Explainable Machine Learning

The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing	97
<i>Andrea Agiollo, Luciano Cavalcante Siebert, Pradeep Kumar Murukannaiah, and Andrea Omicini</i>	
Bottom-Up and Top-Down Workflows for Hypercube-And Clustering-Based Knowledge Extractors	116
<i>Federico Sabbatini and Roberta Calegari</i>	
Imperative Action Masking for Safe Exploration in Reinforcement Learning	130
<i>Sumanta Dey, Sharat Bhat, Pallab Dasgupta, and Soumyajit Dey</i>	
Reinforcement Learning in Cyclic Environmental Changes for Agents in Non-Communicative Environments: A Theoretical Approach	143
<i>Fumito Umano and Keiki Takadama</i>	

Inherently Interpretable Deep Reinforcement Learning Through Online
Mimicking 160
Andreas Kontogiannis and George A. Vouros

Counterfactual, Contrastive, and Hierarchical Explanations
with Contextual Importance and Utility 180
Kary Främling

Cross-Domain Applied XAI

Explanation Generation via Compositional Rules Extraction for Head
and Neck Cancer Classification 187
*Victor Contreras, Andrea Bagante, Niccolò Marini,
Michael Schumacher, Vincent Andrearczyk, and Davide Calvaresi*

Metrics for Evaluating Explainable Recommender Systems 212
Joris Hulstijn, Igor Tchappi, Amro Najjar, and Reyhan Aydoğın

Leveraging Imperfect Explanations for Plan Recognition Problems 231
Ahmad Alelaimat, Aditya Ghose, and Hoa Khanh Dam

Reinterpreting Vulnerability to Tackle Deception in Principles-Based XAI
for Human-Computer Interaction 249
Rachele Carli and Davide Calvaresi

Enhancing Wearable Technologies for Dementia Care: A Cognitive
Architecture Approach 270
*Matija Franklin, David Lagnado, Chulhong Min, Akhil Mathur,
and Fahim Kawsar*

Author Index 281