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.
Explainable and Transparent AI and Multi-Agent Systems

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

Niccolò Marini  
University of Applied Sciences Western Switzerland, Switzerland

Salima Lamisiyah  
University of Luxembourg, Luxembourg

Stephane Galland  
UTBM, France

Giovanni Ciatto  
University of Bologna, Italy
<table>
<thead>
<tr>
<th>Name</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matteo Magnini</td>
<td>University of Bologna, Italy</td>
</tr>
<tr>
<td>Viviana Mascardi</td>
<td>University of Genoa, Italy</td>
</tr>
<tr>
<td>Giovanni Sileno</td>
<td>University of Amsterdam, The Netherlands</td>
</tr>
<tr>
<td>Sarath Sreedharan</td>
<td>Arizona State University, USA</td>
</tr>
</tbody>
</table>
## Contents

### Explainable Agents and Multi-Agent Systems

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

### Explainable Machine Learning

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

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