2025

Prof. Davide Calvaresi, Dr. Amro Najjar, Prof. Andrea Omicini, Prof. Kary Framling

Detroit, 19.05.2025

Morning Session

8:50 – 9:00

Welcome & Workshop Opening

Speaker(s): Prof. Davide Calvaresi, Dr. Amro Najjar, Prof. Kary Framling

9:00 – 09:45

Morning Opening Keynote

Title: Explanations in Responsible Autonomy: Trustworthy AI, Norm Deviation, and Consent

Speaker(s): Prof. Munindar Singh

Abstract: This talk will summarize some of our recent conceptual work on responsible autonomy, highlighting the opportunities and challenges for explanations. This work steps back from computational models of morality to apply insights from philosophy and social psychology to responsible autonomy. One challenge is to understand when responsible autonomy involves respecting of deviating from norms. We show how Habermas's notion of objective, subjective, practical validity criteria can be used to understand norm deviation. We apply case law from the US, UK, and Canada as a source of empirical knowledge about when norm deviations are legitimate. We apply the Habermasian framework as a basis for conceiving of consent. We adapt a model of trust based on the components of ability, benevolence, and integrity to show how trustworthiness involves, again applying case law as a source of empirical intuitions about how to evaluate AI agents. These models can provide a basis for effective and meaningful explanations as integral to responsible autonomy.

09:45 – 10:15

Paper Session 1: Agent Architectures & Inter-Agent Explainability

Session Chair: Prof. Davide Calvaresi

09:45 – 10:00

Title: Ladder of Intentions: unifying agent architectures for explainability and transferability

Authors: Victor Gimenez-Abalos, Adrian Tormos, Filip Edström, Sergio Alvarez-Napagao, Javier Vázquez-Salceda, Mattias Brännström and John Lindqvist

10:00 – 10:15

Title: Engineering Inter-Agent Explainability in BDI Agents

Authors: Katharine Beaumont, Elena Yan, Samuele Burattini and Rem Collier

10:15 – 10:45

Coffee Break (aligned with AAMAS schedule)

10:45 – 12:30

Paper Session 2: Interpretation & Social Dimensions

Session Chair: Dr. Amro Najjar

10:45 – 11:05

Title: Explaining Autonomous Vehicles with Intention-aware Policy Graphs

Authors: Sara Montese, Sergio Alvarez-Napagao, Victor Gimenez-Abalos, Atia Cortés and Ulises Cortés

11:05 – 11:25

Title: Exploiting Constraint Reasoning to Build Graphical Explanations for Mixed-Integer Linear Programming

Authors: Roger Xavier Lera-Leri, Filippo Bistaffa, Athina Georgara and Juan Antonio Rodriguez Aguilar

11:25 – 11:45

Title: Enhancing Surrogate Decision Trees for Reinforcement Learning with Feature Importance Matrices

Authors: Bryan Lavender and Sandip Sen

11:45 – 12:05

Title: Automated Semantic Rules Detection (ASRD) for Emergent Communication Interpretation

Authors: Bastien Vanderplaetse, Xavier Siebert and Stéphane Dupont

12:05 – 12:25

Title: Social Explainable AI: What Is It and How to Make It Happen with CIU?

Authors: Kary Främling

12:30 – 14:00

Lunch Break (aligned with AAMAS schedule)

Afternoon Session

14:00 – 14:45

Afternoon Opening Keynote

Title: Substantial fairness in AI ethics: the path forward

Speaker(s): Prof. Simona Tiribelli

Abstract: Prominent debates in AI ethics contend that biases are one of the main problems to be eradicated to design fair AI systems, leading the community to develop many technical and non-technical methods focused on eliminating them to create fairer AI tools. However, such methods are shown very often to be inadequate or counterproductive for the design of fairer AI systems, failing in fostering social justice as fairness through AI, especially in the domain of healthcare. As I will argue in this talk, one of the main problems is a procedural and merely mathematical understanding of fairness, which leads a passive understanding of biases just as mere errors and technical bugs in AI systems, as well as of the related solutions mainly aimed at their simple removal. To address this problem, this talk will go beyond mathematical and procedural conception of fairness, drawing conceptual insights from moral philosophy, to show what substantial fairness means and how to achieve it by leveraging responsible multi-agent architectures and explanations. Particularly, through a case study on gender and ethnic biases in healthcare AI, this talk will show how a substantial account fairness helps harnessing bias in developing multi-agent architecture-based AI systems to decode patterns of unfairness and, drawing on ethical theories of affirmative action, generate novel compensatory design actions for the development of truly fair healthcare AI systems, capable of addressing longstanding unfair inequalities in healthcare, therefore promoting through AI better and more just healthcare ecosystems..

14:45 – 15:45

Session 3: LLMs & Generative AI in XAI

Session Chair: Prof. Kary Framling

14:45 – 15:05

Title: Evaluating Prompt Engineering Techniques for Accuracy and Confidence Elicitation in Medical LLMs

Authors: Nariman Naderi, Zahra Atf, Peter R Lewis, Aref Mahjoubfar, Seyed Amir Ahmad Safavi-Naini and Ali Soroush

15:05 – 15:25

Title: LLM-based Evaluation Methodology of Explanation Strategies

Authors: Ege Soyarar, Reyhan Aydogan, Berk Buzcu and Davide Calvaresi

15:25 – 15:45

Title: Optimised Timing of Multi-Step Explanations for Multiple Users through Reactive Games

Authors: Akhila Bairy, Martin Fränzle and Maike Schwammberger

15:45 – 16:15

Coffee Break (aligned with AAMAS schedule)

16:15 – 17:55

Session 4: Human-XAI Interaction, Quality & Ethics

Session Chair: Dr. Rachele Carli and Prof. Simona Tiribelli

16:15 – 16:35

Title: Balancing Ethical Persuasion and Explanation Transparency in XAI

Authors: Sukriti Bhattacharya, Rachele Carli, Igor Tchappi, Amro Najjar, and Davide Calvaresi

16:35 – 16:55

Title: Beyond Satisfaction: From Placebic to Actionable Explanations For Enhanced Understandability

Authors: Joe Shymanski, Jacob Brue and Sandip Sen

16:55 – 17:15

Title: Building a Decolonized AI Ethics and Governance Framework

Authors: Sofia Pansoni, Emily Beck and Simona Tiribelli

17:15 – 17:35

Title: Designing for Effective Human-XAI Inter-action

Authors: Mohammad Naiseh, Huseyin Dogan, Stephen Giff, Avleen Malhi and Nan Jiang

17:35 – 17:45

Closing Remarks

A leading venue to discuss explainable agency with cross-disciplinary contributions and top keynote speakers

Prof. Kary Framling
University of Aalto
Explainable ML

Dr. Serena Villata
CNRS
NL, XAI and Argumentation

Prof. Michael Winikoff
University Otago
Explainable Agents

Prof. Bertram Malle
Brown University
Justification to trust HMI

Prof. Dov Gabbai
King's College
XAI and Reasoning

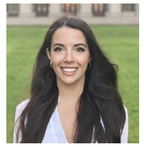
Dr. Julie Shah
MIT
Social/Ethical responsibilities of (X)AI

Prof. Munindar P. Singh
NC State University
XAI in Responsible Autonomy

Prof. Tim Miller
University of Melbourne
XAI and social sciences

Prof. Prof. Brian Lim
University of Singapore
Human-centric XAI

Prof. Prof. Brian Lim
University of Singapore
Enhancing XAI for HIM



Prof. Simona Tiribelli
UniMC & MIT
Substantial fairness in AI ETH

●●●