TUTORIAL INFORMATION

Title	Reusable technologies for explainable autonomous systems
Status (Online or Onsite)	Onsite
Tutors	Davide Calvaresi Giovanni Ciatto Victor Hugo Contreras
Abstract	The opacity of sub-symbolic machine-learning predictors has eventually become unacceptable. To this end, recent studies have focused on providing symbolic knowledge extraction (SKE) and injection (SKI) from and into neural networks, support vector machines, random forests, linear models, etc. By symbolic, we intend intelligible formulations naturally interpretable for both human beings and computers. SKE and SKI are the two main classes of mechanisms serving the intra-agent explainability -- from the data to a numeric prediction/classification, and inter-agent explainability – argumentation-based knowledge alignment among virtual autonomous entities. In particular, the tutorial will focus on: **1)** Introduction to intra-/inter agent explainability and the vision of "EXPECTATION" -- a CHIST-ERA project on XAI. **2)** PsyKE: a platform providing general-purpose support to symbolic knowledge extraction from different sorts of black-box predictors via many extraction algorithms. Notably, PSyKE targets the extraction of symbolic knowledge in logic form, making it possible to extract first-order logic clauses as output. The extracted knowledge is thus both machine- and human- interpretable, and it can be used as a starting point for further symbolic processing—e.g., automated reasoning. **3)** PsyKI: a platform providing general-purpose support to symbolic knowledge injection into predictors via different algorithms. **4)** DEXiRE: a tool for propositional rules extraction from neural networks via binarization

Prerequisites	Basic knowledge of Machine Learning Basic boolean logic knowledge Basic knowledge of Deep learning Basic knowledge of rule-based systems Familiarity with Python and SciKit-learn **Docker installed and working on attendees' PC**
Outline if available	[10 min, Davide] Introduction

Bio of the tutors	**Dr. Davide Calvaresi** is a Senior Researcher at the University of Applied Sciences and Arts Western Switzerland (HES-SO). Davide's research spans over several areas, such as real-time multi-agent systems, explainable AI, blockchain, and assistive technologies. He has been chair of several workshops in those areas, such as EXTRAAMAS (2019-2022), BCT4MAS (2018-2020), and RTcMAS2018. Currently, his main endeavor focus on bridging sub-symbolic and symbolic AI to foster explainability in Multi-Agent Systems in the context of the European Project EXPECTATION. Finally, Davide is also a co-founder of the StartUp Wriggle Solutions pursuing safety on the street with the project SMARTtyre -- real-time monitoring of the tires' consumption and structure -- holding two patents. **Giovanni Ciatto** is a post-doctoral research fellow at the Departement of Computer Science and Engineering (DISI), with a Ph.D. in Data Science and Computation and a MSc in Computer Science and Engineering. Main research interests comprehend the area of artificial intelligence, logic programming, multi-agent systems, and distributed systems. In particular, my contributions have a focus on the models, architectures, and

infrastructures for symbolic reasoning and sub-symbolic processing, there including: (i) knowledge injection and extraction of symbolic knowledge into/from sub-symbolic predictors, (ii) logic and agent-oriented programming, and (iii) the engineering of software technologies laying at their intersection.

Victor Hugo Contreras is a second-year information systems Ph.D. student at the University of Geneva and an e-doctorate at the University of Applied Sciences Western Switzerland (HES-SO), funded by the European project Expectation (Chistera). He received his bachelor's degree in Systems and computer engineering and a master's degree in Systems and computer engineering from the National University of Colombia. His field of research is currently in explainable artificial intelligence XAI, multi-agent programming, chatbot design/implementation, and applied machine and deep learning to multimodal data processing.