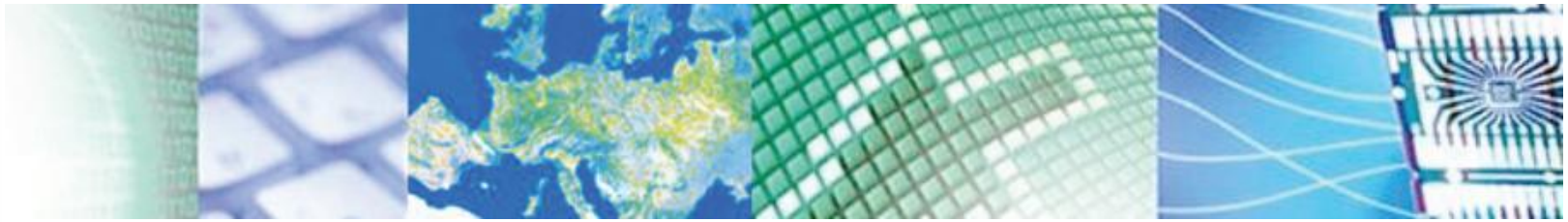


CHIST-ERA Projects Seminar 2021
Explainable Machine Learning-based Artificial Intelligence (XAI)

Davide Calvaresi

April 14, 2021

Programme co-funded by the EUROPEAN UNION

Introduction: Application Scenarios of the Topic

EXPECTATION - Food and **nutrition recommender** systems

CIMPLE - Misinformation, **information manipulation** (climate change and covid), creative response

INFORM - Radiomics in Oncology and **Medical Imaging**

COHERENT - User interactions in **assistive robotic** manipulation tasks

CausalXRL - intensive care, neuro-rehab, neuromorphic implementation, farming, e-learning **(multi-dom)**

XAIface - face-recognition for pedestrians **flow management**

MUCCA - basic science (high energy physics), **medicine** (med & functional imaging), **neuro-science** (brain enc of complex beh)

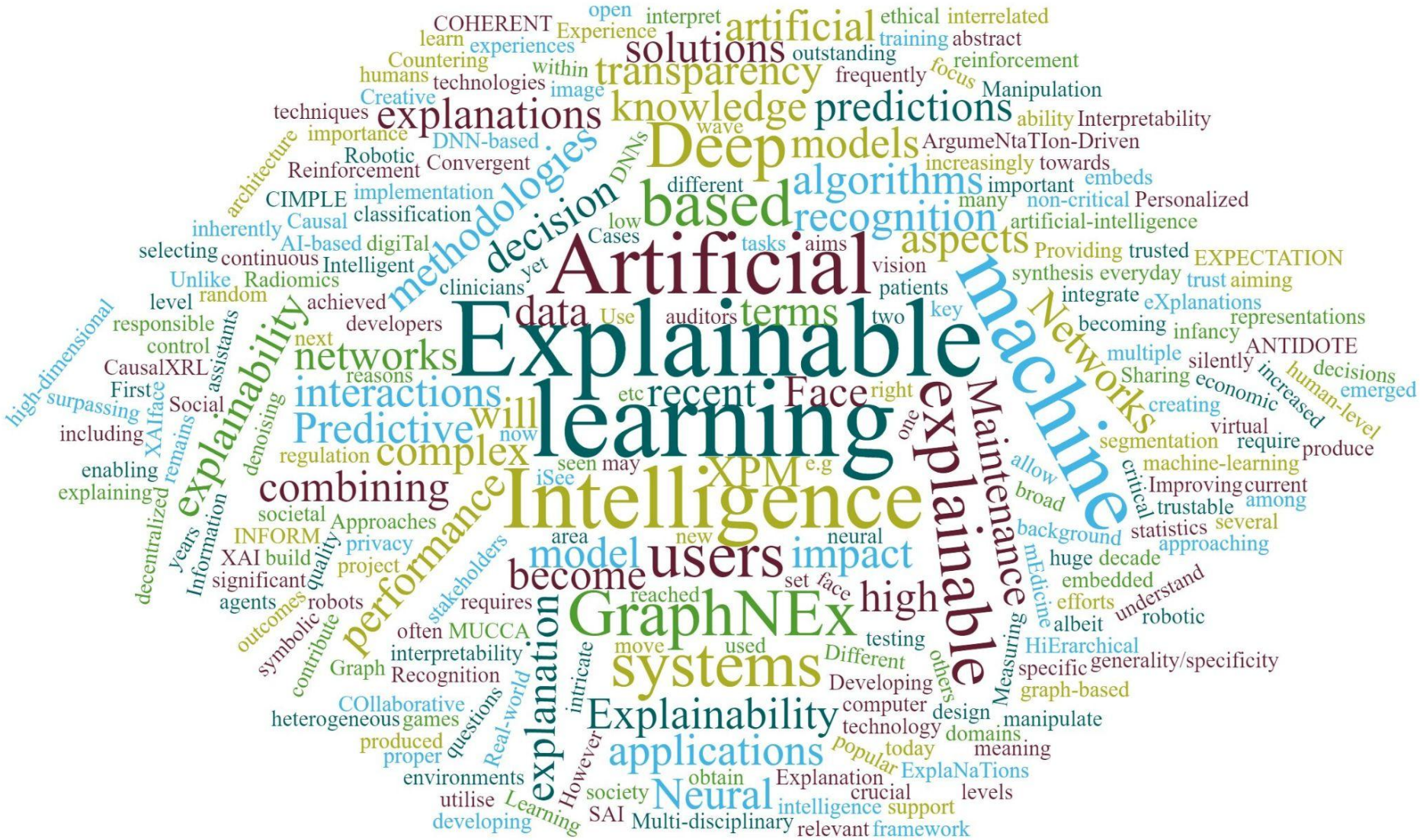
iSee - Telecom, Medical Radiology, Natural Environmental Event Detection, **Cyber security**

ANTIDOTE - digital **medicine**

XPM - electric vehicles, metro trains, steel plants, wind farms **(transportation/multi-dom)**

GraphNex - genomics, privacy

Introduction: Projects of the Topic



Introduction: Projects of the Topic

CIMPLE - Info manipulation detection, **Interpretable NLP** with **graphs, creative explanations** and **viz**
EXPECTATION - Multimodal, Personalized, & **Distributed** explanations, **Negotiation**, **Heterogeneity**
CausalXRL - Explainable decision support, **Causal inference**, **Model-based** Reinforcement Learning
XPM - Explainable Predictive Maintenance, **Contextualization**, Understandability, **Multi-actor**
iSee - Sharing, Reusing, and evaluating XAI, **Human-centric**, and **Personalized** explanations
COHERENT - Explanations **Adaption/Personalization**, **Combination**, and **Reconciliation**
MUCCA - Feedback-based XAI, general procedures, and engineering pipelines
ANTIDOTE - Argumentation-driven explanations and **Interpretable NLP**
XAIface - Security, **Trust**, Social Acceptance, Fairness in face recognition
GraphNex - graphs, **heterogeneous data**, **user interaction**
INFORM - DNN Interpretability to enforce **Trust**
SAI - Distributed and **Human-centric**

Expected Achievements and Outputs

Build **Trust** in artificial intelligence [ALL]
Building blocks (**Toolbox**) for explainable AI-"variants" [iSee, XAIface, CausalXRL, EXPECTATION]
Sets of explanation **strategies** [iSee, MUCCA, XPM, CIMPLE, EXPECTATION]
Novel post hoc explainability layers for black-box AI [XPM, MUCCA]
Novel inherently interpretable/explainable AI models [XPM, INFORM, SAI, GraphNex, CausalXRL, XAIface, CIMPLE, ANTIDOTE, EXPECTATION]
XAI at different levels/different users (**Personalization**) [iSee, SAI, CausalXRL, COHERENT,GraphNex, CIMPLE, MUCCA, EXPECTATION]
Decision making/support algorithms [ANTIDOTE, INFORM, CausalXRL, XPM]
Enable **decentralised** XAI [EXPECTATION, SAI]
User-centered full **control on data** and AI models they share [SAI]
Incorporate **human-behavioral** models in XAI [EXPECTATION, SAI, iSee]
Propose multi-faceted evaluation **metrics** for explanations [iSee, XPM, SAI, XAIface, COHERENT]
Proof of concept XAI on particular applications [ALL]
Cross-application assessment of available approaches to XAI (**generalization**) [iSee, MUCCA, EXPECTATION]
Human-centered **Interactive explanations** [EXPECTATION, iSee, XPM, ANTIDOTE, CIMPLE]
Store, reuse or **combine** explanation experiences [iSee, COHERENT]
Open Source **platform** and European XAI **compliance certification** framework [iSee]
XAI **Social Networks** [iSee]

Upcoming Challenges and Needs

Interpretability to explainability (**subsymbolic to symbolic**) Generalization [EXPECTATION, CausalXRL, CIMPLE]
Common representation for **heterogeneous** explanation methods [iSee, EXPECTATION]
Heterogeneous **explanations reconciliation** (i.e., diverging exp) [EXPECTATION]
User-tailored **explanation manipulation** [COHERENT, EXPECTATION, CIMPLE, iSee]
Adaptability to various data structures, models architectures, & tasks [MUCCA, COHERENT, INFORM,GraphNex, CausalXRL]
Unify existing libraries of existing XAI techniques [iSee, COHERENT]
Evaluate **post-hoc** fashion vs. build **inherently** interpretable models [XPM, COHERENT]
Differentiate explanations to support **different actors** [EXPECTATION, XPM, COHERENT, ANTIDOTE, SAI, CIMPLE]
Evaluation of explanations [iSee, COHERENT, MUCCA, XPM, INFORM, XAIface, ANTIDOTE, SAI, CIMPLE]
Define the **human-centric elements** to be incorporated [SAI, CIMPLE]
Identify latent states & actions in hierarchical causal models inferred from data human-interpretable [CausalXRL]
Combine a **task-specific** prediction models and a **general NL** for explanatory dialogues [ANTIDOTE, EXPECTATION]

XAI **community** engagement, collaborations and standardization [iSee, XAIface, COHERENT]

Possible Roadmap

Define/agree on **criteria/metrics** for "**correctness**", "**successful**", and "**quality**" explanations [All]

Testing paradigm definition and design [EXPECTATION, MUCCA, CausalXRL]
Ensuring **data-availability** + GDPR compliance [EXPECTATION, XAIface,ANTIDOTE, MUCCA, iSee]
Participatory **stakeholder engagement** events (co-creation / co-design) [XAIface, iSee, CIMPLE]
Ethical involvement/assessment (contribution to defining AI regulations) [XAIface,iSee, EXPECTATION]

User-profiling for EXP personalization [EXPECTATION, COHERENT, iSee]
Identify **human-centric models/elements** for XAI [SAI, iSee, CIMPLE]
Psychology/human perception of explanation design [iSee,CIMPLE, COHERENT]

Injection of **symbolic** knowledge **into subsymbolic** predictors [EXPECTATION, CIMPLE, CausalXRL]
Identify existing AI models suitable for **decentralization** [EXPECTATION, SAI]
Explanation misalignment resolution [EXPECTATION]
Share explanation experiences using standard **vocabularies** [iSee]
Ontology and **Knowledge alignment** [EXPECTATION, COHERENT, iSee, GraphNex, CIMPLE, CausalXRL]
To **benchmark** task-specific XAI leveraging external knowledge (models & simulation) [MUCCA, COHERENT]
…
Industry engagement and differentiating AI design users versus end-users [iSee]

…

Role of the CHIST-ERA Support

Good:

Early **support** already provided by CHIST-ERA
Clear **guidelines**
Prompt **replies**
OpenAIRE CHIST-ERA Course on **Open Science** for funded projects
Current CHIST-ERA **seminar**.
Opportunities to find collaborators and **create synergies** within the several consortia

Difficulty:

Different National regulations / Synchronization of national regulations

Responsible Research & Innovation

Besides the measures already in place…

Liability definition for explainable intelligent systems
Leverage the **support of existing European initiatives** on responsible & ethical research (e.g., SoBigData++)

Promotion of **public engagement** to build **usable** XAI methods
Gender **balance**: it's improving, but we are not there yet.

Projects outcomes should contribute to **national** and **international** governmental policy, by:
Creating discourse **around GDPR** (increasing awareness)
Certification of tried-and-tested XAI methods

Increasingly promoted Open-access: Publications, datasets, reproducible code (through open repositories)
Open-access processes needs **more clarity** (waive/not-waive) and **cut their costs!**
Accountability: data management plans, project progress reports

Open Science

Old but gold

Published works will be publicly available by targeting journals and conferences providing **free-online access** to the papers and by providing open access to the author versions of the articles on the partners' websites as well as on publicly accessible article repositories such as arXiv.org and OpenAIRE

Open source **development**

When possible, **datasets** and **methods** will be **shared** on existing platforms (e.g., SoBigData++, AI4EU, XAIface)

Data Management Plan (DMP) **to ensure the availability/accessibility** of the produced data

New opportunities

Towards the end of this course, possibly **redacting a XAI text book** (or a research volume/special issue)

In the context of EXTRAAMAS 2022 (https://extraamas.ehealth.hevs.ch):

➔ Special Issue (OA) and dedicated track for CHISTERA XAI ongoing works

➔ Gender-balanced panel for XAI's young researchers (publicly accessible)

Next year resolution: what worked and what did not

Technology Transfer

Integration of novel solutions into existing commercial solutions [EXPECTATION, XAIface, iSee]
An improved decision-making process for industrial **maintenance** currently BB [XPM]
Increased **awareness** about **pros/cons** of glass models vs. BB [XPM]
Market analysis and potentiality of novel software tools (AI-based) [INFORM]
Cross-domain knowledge transfer [iSee, EXPECTATION, MUCCA, XPM, INFORM, CausalXRL]
Transfer initial results to **national research centers** [SAI]
Adoption of XAI in **healthcare** (Med schools and health institutes) [ANTIDOTE, MUCCA, INFORM,EXPECTATION, CausalXRL, iSee]
Technology transfer in **security** [XAIface, iSee]
Adoption of XAI in **media** [CIMPLE]
Adoption of AI technologies in basic (**theoretical**) research [EXPECTATION, MUCCA]

Training activities (e.g., training schools) [All]
Establishing **community** and **social network** of XAI researchers [All]
Create a **European XAI compliance certification** for AI software [All]

Questions

Thanks...Questions?

CausalXRL **CIMPLE**

inform

XPM

iSee
www.isee4xai.com

XPECTATION
PERSONALIZED EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR DECENTRALIZED AGENTS WITH HETEROGENEOUS KNOWLEDGE

MUCCA

COHERENT

SAI
SOCIAL EXPLAINABLE AI