The AAAI's Workshop on
Artificial Intelligence Safety

EnnCore: End-to-End Conceptual Guarding of Neural Architectures*

Edoardo Manino,[1] Danilo Carvalho,[1] Yi Dong,[2] Julia Rozanova,[1] Xidan Song,[1] Mustafa A. Mustafa,[1,3] Andre Freitas,[1,4] Gavin Brown,[1] Mikel Luján,[1] Xiaowei Huang,[2] and Lucas Cordeiro[1]

[1]Department of Computer Science, The University of Manchester
[2]Department of Computer Science, University of Liverpool
[3]imec-COSIC, KU Leuven,
[4]Idiap Research Institute

*EPSRC Reference: EP/T026995/1
https://enncore.github.io/

Vision and Challenges

Build **explainable** and **fully-verifiable learning-based systems** that are **safe, transparent** and **robust**

- Trade-off between **soundness** and **completeness** to achieve **scalability**
- **Representational gap** between the **neural** (flexibility) and the **symbolic** (explainability, control)
- Trade-off between **privacy protection** and **transparency / accountability**

Objective

Develop and evaluate methods, algorithms and tools to build explainable and fully-verifiable learning-based systems that are safe, transparent and robust

Objective of this special session

- Establish new **partnerships/collaborations** to

  (1) lead the discussion concerning **the challenges and opportunities**
  (2) tackle our main challenges to achieve **explainable and fully-verifiable learning-based systems**
  (3) create a **benefits roadmap** in collaboration with the SafeAI community

- Contact lucas.cordeiro@manchester.ac.uk if you are interested in collaborating with us

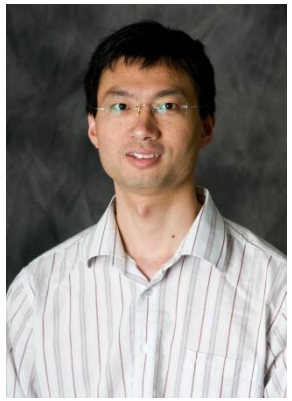
EnnCore

EnnCore Team

L. Cordeiro

A. Freitas

M. Mustafa



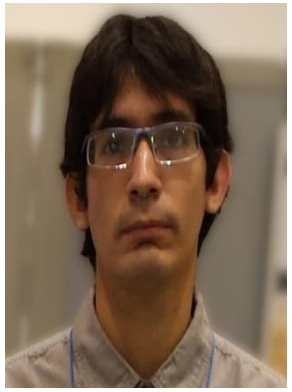
X. Huang

G. Brown

M. Luján

E. Manino

Y. Dong



D. Carvalho

J. Rozanova

RA in Secure & Privacy-Preserving AI Models

EnnCore

EnnCore Partners

MANCHESTER 1824
The University of Manchester

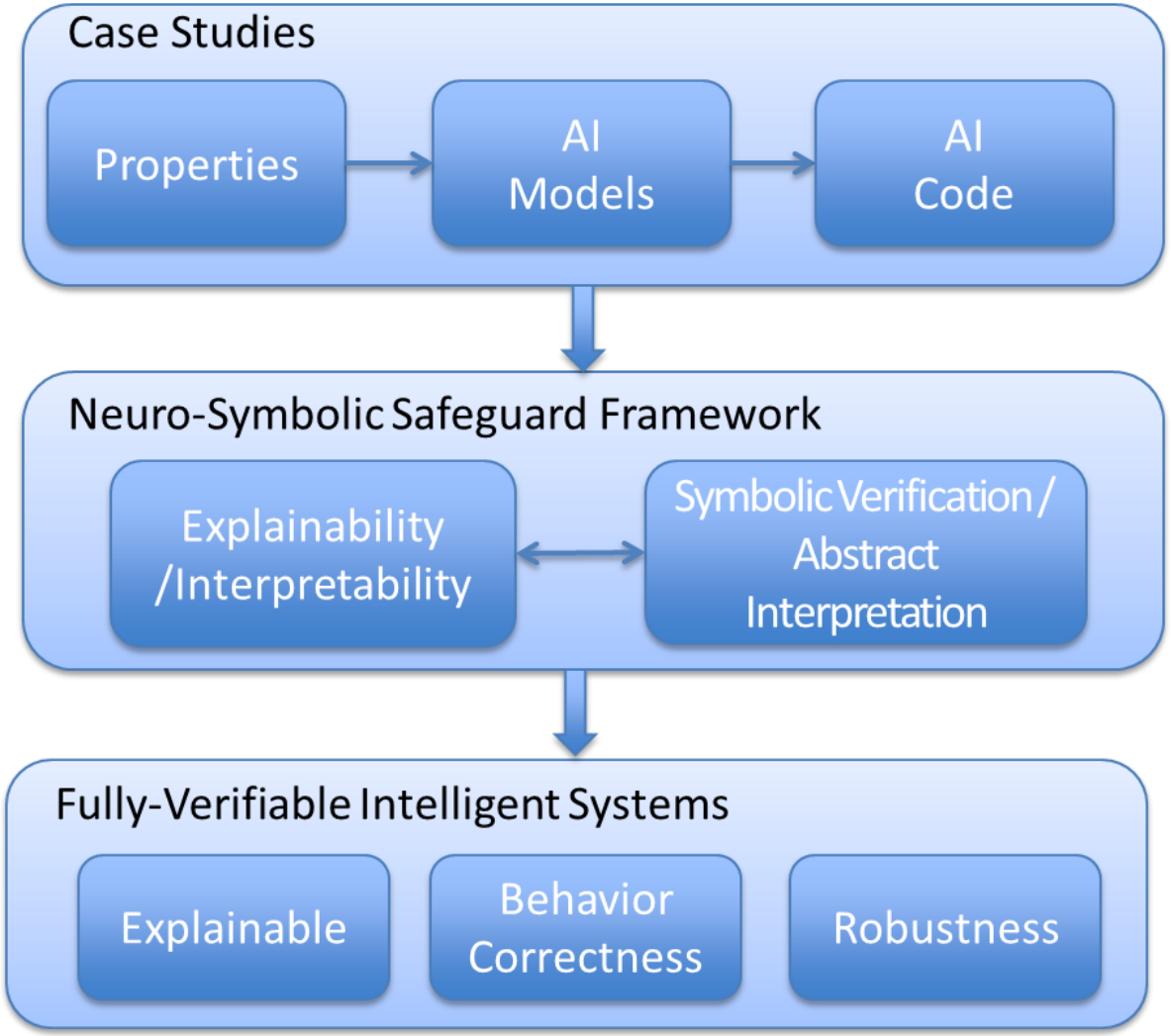
UNIVERSITY OF LIVERPOOL

URBANCHAIN

CANCER RESEARCH UK

EPSRC
Engineering and Physical Sciences Research Council

EnnCore: Proposed Research

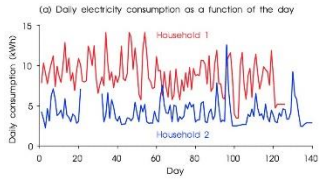


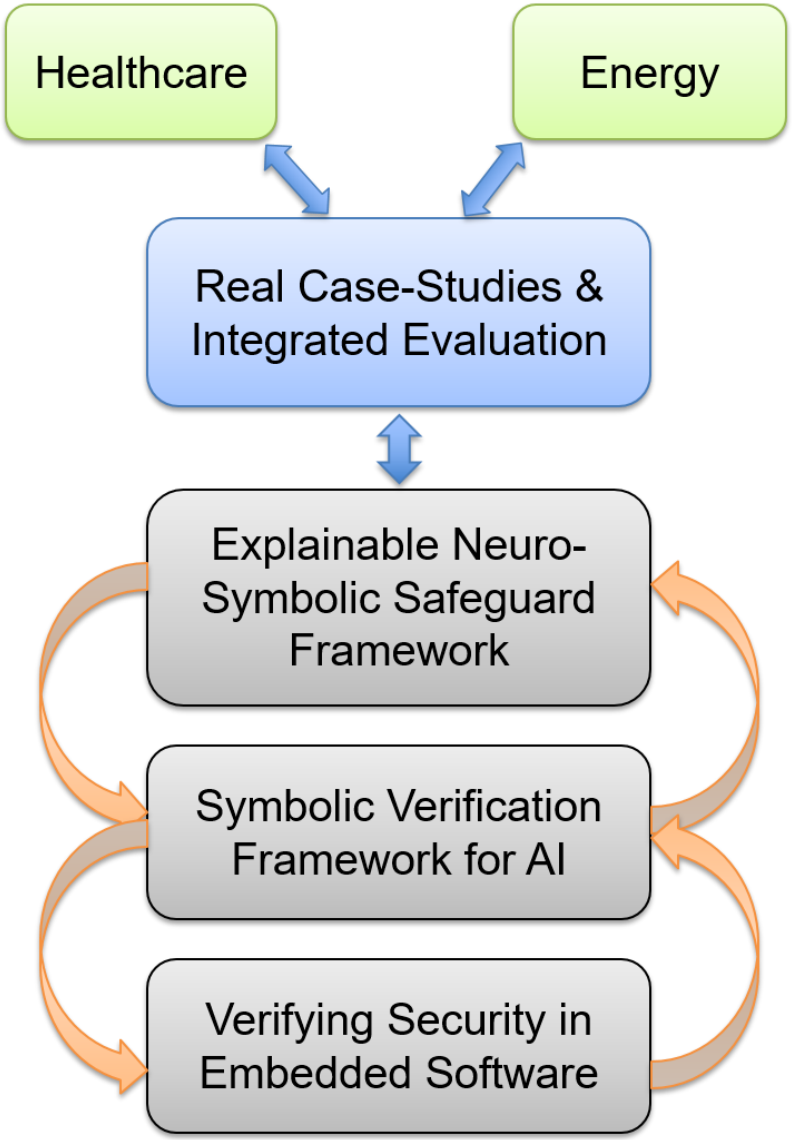
- Creation of the evaluation benchmarks
- Use case deployment & usability study

- Develop neural interpretability methods
- Reason over security properties in DNN implementations

- Evaluation of security properties in real case studies
- Validation of the results

EnnCore





EnnCore Talks

Privacy Friendly Energy Consumption Prediction: Real Case-Studies, Mustafa A. Mustafa (University of Manchester, UK / KU Leuven, Belgium)

Explainability and Inference Controls, André Freitas (University of Manchester UK & Idiap Research Institute, Switzerland)

Safety Verification of Deep Reinforcement Learning, Yi Dong (University of Liverpool, UK)

Verifying Quantized Neural Networks using SMT-Based Model Checking, Edoardo Manino (University of Manchester, UK)

EnnCore

EnnCore Website

(https://enncore.github.io/)

EnnCore

End-to-End Conceptual Guarding of Neural Architectures

- News
- Partners
- Positions
- Publications
- People
- Applications
- Third Party Contributions
- Index of Benchmarks
- Awards

Partners

MANCHESTER 1824
The University of Manchester

UNIVERSITY OF LIVERPOOL

EPSRC
Engineering and Physical Sciences Research Council

CANCER RESEARCH UK

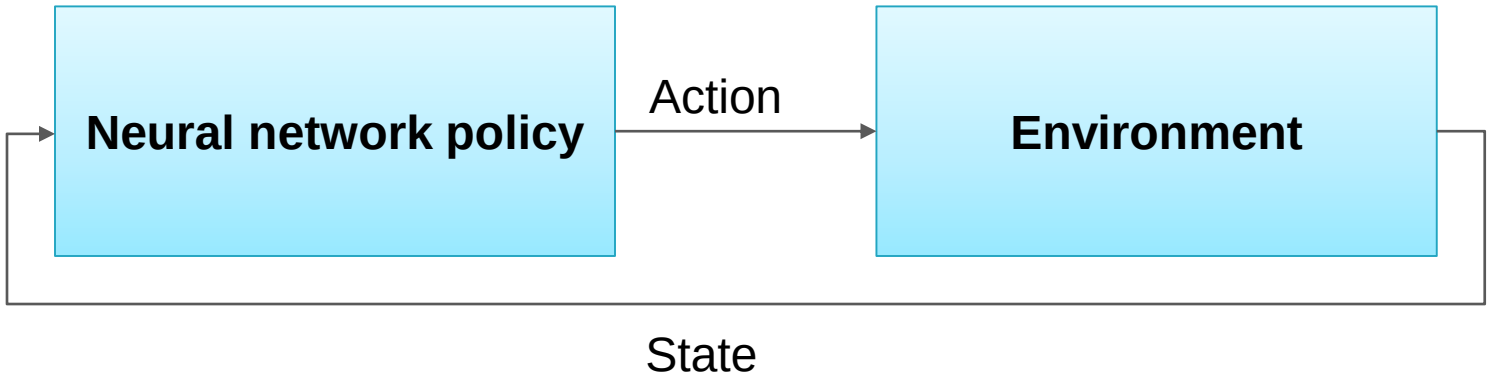
URBANCHAIN

Invited Talk

(Mathias Lechner from IST Austria)

Closed-loop Safety of Bayesian Neural Networks and Stochastic Control Systems

- Discusses recent works on verifying the safety of closed-loop stochastic systems with neural network control policies



Safety: System never reaches unsafe states

Stability: System always reaches target states

EnnCore

EnnCore Program at SafeAI 2022

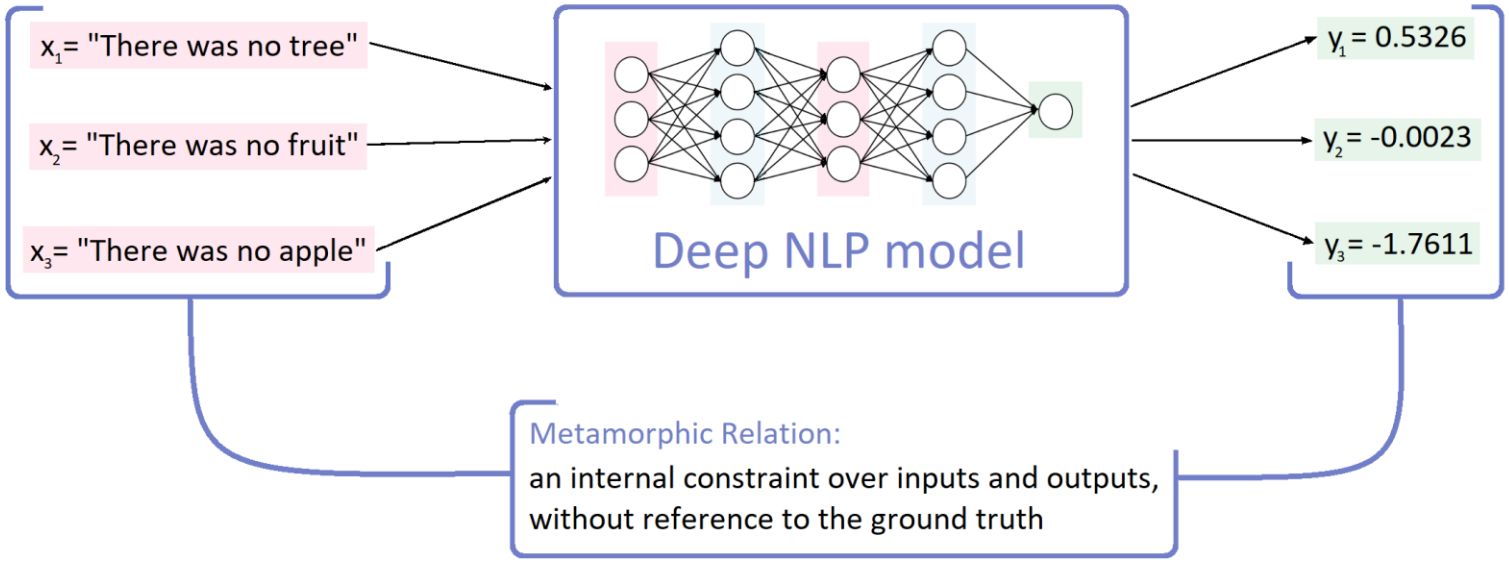
(https://safeai.webs.upv.es/)

Time (UTC)	Description
14:00-14:10	Welcome, overview, Lucas Cordeiro (University of Manchester, UK)
14:10-14:25	Verifying Quantized Neural Networks using SMT-Based Model Checking, Edoardo Manino (University of Manchester, UK)
14:25-14:40	Explainability and Inference Controls, André Freitas (University of Manchester UK & Idiap Research Institute, Switzerland)
14:40-14:55	Safety Verification of Deep Reinforcement Learning, Yi Dong (University of Liverpool, UK)
14:55-15:10	Privacy Friendly Energy Consumption Prediction: Real Case-Studies, Mustafa A. Mustafa (University of Manchester, UK / KU Leuven, Belgium)
15:10-15:30	Closed-loop Safety of Bayesian Neural Networks and Stochastic Control Systems, Mathias Lechner, IST Austria

Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective

Edoardo Manino, Julia Rozanova, Danilo Carvalho, André Freitas, Lucas Cordeiro

60th Annual Meeting of the Association for Computational Linguistics



Challenge:

deep neural networks are black boxes. Specifying our expectations on their internal behaviour is not trivial

Contribution:

we express systematicity, compositionality and transitivity as metamorphic relations. Thanks to this, we can test the internal linguistic consistency of state-of-the-art NLP models

MANCHESTER 1824
The University of Manchester

Full Paper:
https://openreview.net/forum?id=Lxf2vB1YTG2

Website:
https://enncore.github.io/

EnnCore

EnnCore