TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR
Grant Agreement Number 952215

D3.4 Handbook on Trustworthy AI Report (v2)

Document type (nature)	Report
Deliverable No	3.4
Work package number(s)	3
Date	Due M46, 30 June 2024
Responsible Beneficiary	CNR, ID #2
Author(s)	Umberto Straccia (CNR), Francesca Pratesi (CNR) For contributors, see the related section.
Publicity level	Public
Short description	Handbook on Trustworthy AI

History			
Revision	**Date**	**Modification**	**Author**
1.0	2024-07-04	Version 2	Umberto Straccia

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Fredrik Heintz	1 / LiU	2024-07-08
Michela Milano	10 / UNIBO	2024-07-07

Table of Contents

Summary of the report

WP3 decided to write the deliverable encyclopaedia-like and present it in the form of a publically accessible WIKI. To do so, the Jupiter Book framework has been used.

This report describes the structure and design of the TAILOR Handbook on Trustworthy AI wiki, which can be accessed from

https://tailor-network.eu/handbook/

and is physically hosted at

http://tailor.isti.cnr.it/handbookTAI/TAILOR.html

An automatically generated pdf of the wiki is appended to this document.

Introduction

The Handbook on Trustworthy AI assumes an encyclopaedia-like structure and is presented in the form of a publicly accessible WIKI. To do so, the Jupiter Book framework has been used.

In the long term, the handbook is meant to become a point of reference for resources (EU| AI act, key concepts, tools, documentation, tutorials, teaching material, etc.) related to Trustworthy AI and the EU AI Act on Trustworthy AI.

Each task leader of WP3 contributed to it and will update the content he is responsible for, as soon as major changes occur during the life period of the project.

An automatically generated pdf of the wiki is appended to this document.

This report describes the current structure and design of the TAILOR Handbook on Trustworthy AI wiki, which can be accessed from

https://tailor-network.eu/handbook/

and is physically hosted at

http://tailor.isti.cnr.it/handbookTAI/TAILOR.html

Contributors

Coordinators:

- Umberto Straccia - Institute of Information Science and Technologies "A. Faedo" of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Francesca Pratesi - Institute of Information Science and Technologies "A. Faedo" of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy

The following additional people have contributed to the Handbook on Trustworthy AI (in alphabetical order):

- Riccardo Albertoni - Istituto di Matematica Applicata e Tecnologie Informatiche "Enrico Magenes", Consiglio Nazionale delle Ricerche (IMATI-CNR), Via De Marini, 6, 16149 Genova, Italy
- Tristan Allard - University of Rennes, CNRS, IRISA, 35000 Rennes, France
- Guilherme Alves, University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
- George Ashwin - Faculty of Electrical Engineering Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
- Alejandra Bringas Colmenarejo, School of Law, University of Southampton, SO17 1BJ, United Kingdom
- Stefan Buijsman - Delft University of Technology, Jaffalaan 5, 2628 BX, Delft, The Netherlands
- Pablo A M Casares - VRAIN, Universitat Politècnica de València
- Sara Colantonio - Institute of Information Science and Technologies "A. Faedo" of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Miguel Couceiro - Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
- Santiago Escobar - VRAIN, Universitat Politècnica de València
- Alessandro Fabris, MPI-SP, Max Planck Institute for Security and Privacy, 44799 Bochum, Germany
- Peter Flach, University of Bristol, United Kingdom
- Gabriel Gonzalez-Castañé - University College Cork, Cork, Ireland
- Riccardo Guidotti - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Fredrik Heintz - Linköping University, Department of Computer and Information Sciences, 58 183 Linköping, Sweden
- Jose Hernandez-Orallo - VRAIN, Universitat Politècnica de València
- Sietze Kuilman - Faculty of Electrical Engineering Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
- Karima Makhlouf, Inria, Ecole Polytechnique, IPP, 91120, Paris, France
- Marta Marchiori Manerba - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Fernando Martinez-Plumed - VRAIN, Universitat Politècnica de València
- Anna Monreale - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Roberto Pellungrini - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Miquel Perello Nieto, University of Bristol, United Kingdom
- Francesca Pratesi - Institute of Information Science and Technologies "A. Faedo" of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Resmi Ramachandran Pillai - Linköping University, Department of Computer and Information Sciences, 58 183 Linköping, Sweden
- Nicola Rossberg - School of Computer Science & IT, University College Cork, Cork, Ireland
- Andrea Rossi - SFI Centre for Research Training in Artificial Intelligence, University College Cork
- Marie-Christine Rousset - University of Grenoble Alpes, Grenoble, France

- Salvatore Ruggieri - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Luciano C Siebert - Faculty of Electrical Engineering Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
- Piotr Skrzypczyński - Institute of Robotics and Machine Intelligence, Poznań University of Technology, ul. Piotrowo 3A, 60-965 Poznań, Poland
- Kacper Sokol, ETH Zurich, Switzerland
- Jerzy Stefanowski - Institute of Computing Science, Poznań University of Technology, ul. Piotrowo 2, 60-965 Poznań, Poland
- Barry O'Sullivan - School of Computer Science & IT, University College Cork, Cork, Ireland
- Andrea Visentin - School of Computer Science & IT, University College Cork, Cork, Ireland
- Arkady Zgonnikov - Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, The Netherlands
- Sami Zhioua, Inria, Ecole Polytechnique, IPP, 91120, Paris, France

Hosting

Currently, the TAILOR Handbook on Trustworthy AI can be accessed from

https://tailor-network.eu/handbook/

and is physically hosted at

http://tailor.isti.cnr.it/handbookTAI/TAILOR.html

Structure of the TAILOR Handbook on Trustworthy AI (v2)

Overall, the handbook content's structure has been inspired by similar encyclopaedia-like works such as the

Encyclopaedia of Machine Learning and Data Mining. Editors: Claude Sammut, Geoffrey I. Webb, 2017. Springer. https://doi.org/10.1007/978-1-4899-7687-1

The major upgrade with respect to Version 1 is that, apart from updating the scientific/technological aspects, the handbook also now includes the salient concepts and terms of the EU AI Act and relates them to the terminology and techniques of the scientific part. So, in summary a reader may get into the main notions the EU AI Act is about from a legal point of view, and then refer to the current techniques available to implement the various different obligations AI systems have to comply with this law.

We have also updated the terminology of the scientific part in order to comply with the one used within the EU AI Act and to mirror the structure of the "Ethics Guidelines for Trustworthy Artificial Intelligence" written by the High-Level Expert Group on AI.

The introductory part is about the Ethical and Legal Framework the handbook is referring to within the European context. We also include a definition of Trustworthy AI and the

European Legal Framework, starting from the Ethical Guidelines for Trustworthy AI and ending with the EU AI Act. We summarise the classification of AI systems based on their level of risk and describe the different obligations AI systems must comply with this law.

The content of the above-mentioned material is described in the following sections of the handbook:

- The TAILOR Handbook of Trustworthy AI (entry point of the handbook)
- Trustworthy AI
- The Ethical and Legal Framework
- Human Agency and Oversight

Besides the above-mentioned material, the Handbook covers then the following dimension of Trustworthy AI

- Transparency
- Technical Robustness and Safety
- Diversity, Non-Discrimination, and Fairness
- Accountability, Reproducibility, and Traceability
- Privacy and Data Governance
- Societal and Environmental Wellbeing

Each dimension has a series of entries associated.

Each entry has:

- Potential synonyms
- Brief summary
- A more detailed section
- Bibliography
- List of authors

There is also an

- Index

that lists all entries in alphabetical order, references to a short definition of an entry and where it is used within the handbook. Potential synonyms have their own entries in this index.

The second version of the Handbook has 30 new entries, for a total number of 93 entries.

Appendix: PDF of the Handbook on Trustworthy AI (v2)

An automatically generated pdf of the handbook's wiki is included in here for completeness. The pdf may not necessarily respect the formatting of the online version (e.g, the formatting related to latex formulae on the pdf may be incorrect). However, the published version is the updated version.