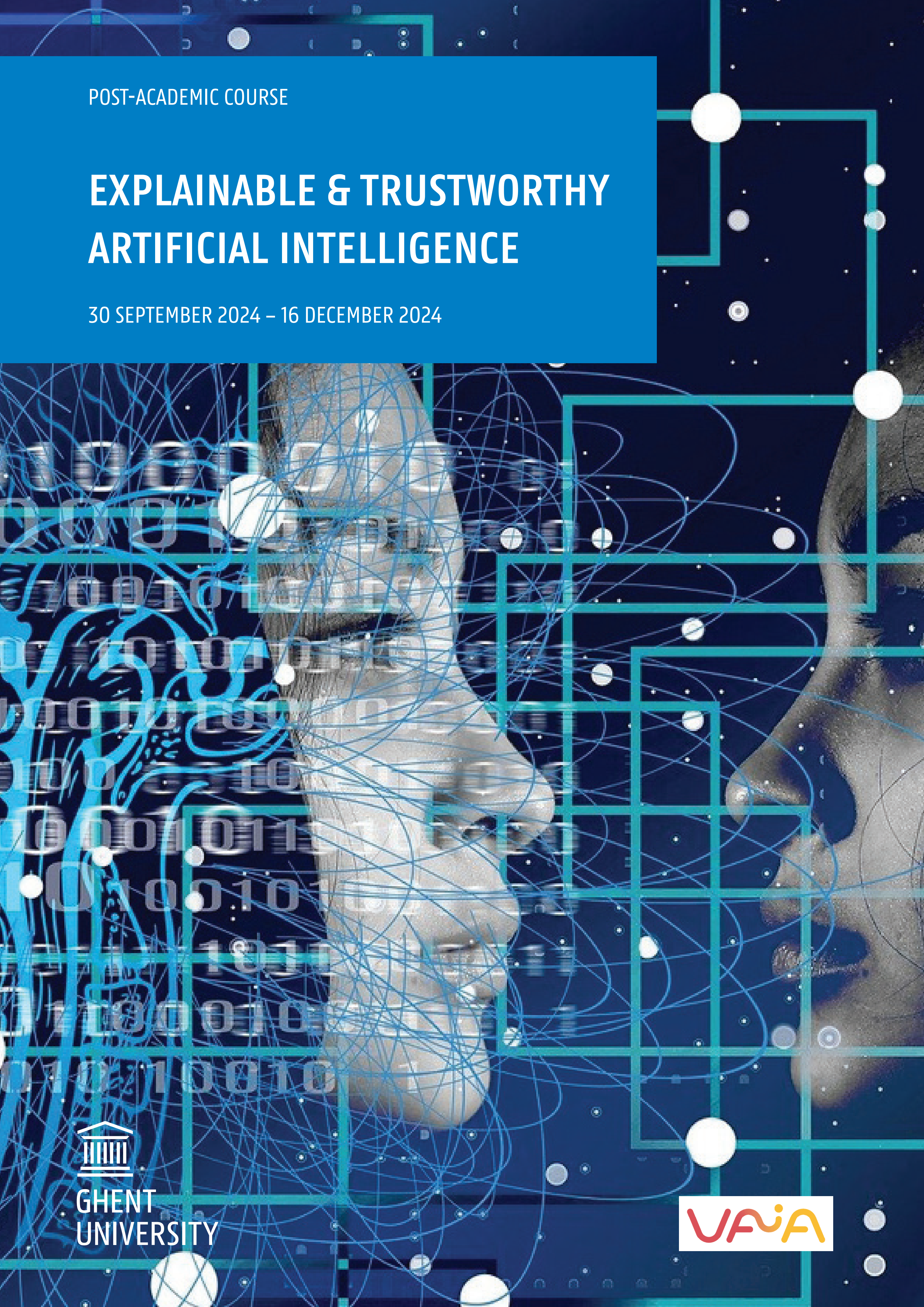
POST-ACADEMIC COURSE

EXPLAINABLE & TRUSTWORTHY ARTIFICIAL INTELLIGENCE

30 SEPTEMBER 2024 – 16 DECEMBER 2024



GHENT
UNIVERSITY

VAIA

Artificial Intelligence (AI) has come a long way since its first use and application many decades ago.

The use of AI and Machine Learning have seen an immense uptake in the 21st century. The techniques developed in the domain were and are **successfully applied to a wide variety of problems,** both in academia, private and public industry. As this domain became more and more established in recent years, **new challenges** arose.

Artificial Intelligence nowadays are complex and sophisticated algorithms that sometimes make it **difficult for the humans to understand and interpret the decisions or suggestions** of the AI system. Explainable AI puts the following properties on the foreground **to deliver trust:**

- Gaining trust by **explaining** for example **the characteristics of AI output.**

- By explaining an AI technique understanding will increase, allowing to **investigate if the technique can be transferred to another domain or problem.**

- **Informing a user about the workings of an AI model** so that there is no misinterpretation.

- Confidence of users can be established by **using AI models that are explainable, stable but also robust.**

- When explaining AI models **issues concerning privacy** awareness come into play. Private data should not be exposed by the models.

- It is important that **actions can be explained.** How have we come to specific outcomes and how could we change them?

- Nowadays, **a wide variety of people from different background** come into contact AI, it is important that **they all understand why the system is behaving in such a manner** and offer **explanations tailored to their needs.**

INFO AND REGISTRATION

WWW.UGAIN.UGENT.BE/EXPLAINABLEAI

WHO SHOULD ATTEND?

The lessons are intended for anyone who has a **good professional familiarity with computer science** and who would like **to get more insights in techniques that can be applied to achieve explainable and trustworthy Artificial Intelligence.** Participants have completed a higher education in computer science or have acquired an equivalent experience.
Participants have programming experience with Python or a related programming language.
The lessons can be followed onsite in the UGain classroom or online (live and on demand).

CERTIFICATE

To receive a certificate, one should attend all the lessons and succeed for the final exam.
The exam will take place on January 27th, 2025.

SCIENTIFIC COORDINATION

prof. dr. Femke De Backere, Department of Information Technology, Ghent University – imec & VAIA

TEACHERS

- **Tijl De Bie,** Department of Electronics and Information Systems, Ghent University
- **Matthias Feys,** ML6
- **Femke Ongenae,** Department of Information Technology, Ghent University
- **Daniel Peralta Cámara,** Department of Information Technology, Ghent University
- **Jonathan Peck,** Department of Applied Mathematics, Computer Science & Statistics, Ghent University
- **Yvan Saeys,** Department of Applied Mathematics, Computer Science & Statistics, Ghent University
- **Sofie Van Hoecke,** Department of Electronics and Information Systems, Ghent University
- **Willem Waegeman,** Department of Data Analysis and Mathematical Modelling, Ghent University

PROGRAMME

1. INTRODUCTION

In this first lesson, we give **a short recap of the basics, followed by the explanation of some general terms** that are used in the domain of explainable and trustworthy Artificial Intelligence.
This introduction will end with the **definition of the challenges** within this domain.

2. WHITE BOX MODELS

While **black box models offer higher accuracy, white box models are easier to explain and to interpret**, unfortunately this **leads to a lesser predictive capacity.**
In the area of white box models, several different approaches will be highlighted: **Linear regression, Generalized Additive Models (GAMs), decision trees and rule-based systems and fuzzy logic.**

3. INTERPRETABILITY & EXPLAINABILITY

Machine learning systems build **models that learn to automate complex tasks by learning from examples**. How to get **insights** into how these models work depends on **the type of algorithm used**.   Getting insights into how our models work can be done by looking at how the model works in general (interpretability), versus how a specific prediction of the model was computed (explainability). Additional hypothetical "What if" questions can be asked to allow for counterfactual reasoning, adding to the toolkit of explainability methods.

4. ONLINE & TRANSFER LEARNING

Training machine learning systems can be done **before use**, i.e. when training it on a stack of pictures first and asking it to make sense of new pictures later. However, it can also be done **during use.** In the latter scenario the system gets updated whilst it is being used. Sometimes this is necessary because training data is (partially) becoming available after commissioning of the system. Sometimes a system is pretrained on one dataset and the developer wants to retrain the system in order to solve another but related problem, i.e. using a machine vision system that is trained to detect cats to now detect dogs. The developer thus leverages the effort put into the training of the earlier system, hence requiring less training time for the novel system. **These and other relations between datasets, their application in training models and the problems we solve with those will be explained in this lesson.**

5. HYBRID AI

The oldest forms of machine learning entail rule engines that were **hand programmed.** Newer forms entail **algorithms searching for connections themselves.** The first are great in explaining how they reach their conclusions. The latter sometimes give superior predictions, being a lot less brittle, but lack that explainability. **To get the best of both worlds, these approaches are sometimes combined.** Moreover, allowing an expert to guide a machine learning system can sometimes lead to yet again superior predictions.

6. ROBUSTNESS

The output of a machine learning system depends on the data used as input. Often the needed amount and structure of that data is overlooked. However, machine learning systems can be combined to generate additional data or to finetune each other. Nevertheless, **malicious additions to your training data can corrupt your system** and even **a well-trained system can be deceived. This lesson explains these issues and what you can do about them**.

7. UNCERTAINTY

The notion of **uncertainty** is of major importance in machine learning and constitutes **a key element** of modern machine learning methodology. In recent years, it has gained attention due to the increasing relevance of machine learning **for practical applications,** many of which are coming with safety requirements. In this regard, **new problems and challenges** have been identified by machine learning scholars, many of which call for novel methodological developments. Indeed, while uncertainty has a long tradition in statistics, and many useful concepts for representing and quantifying uncertainty have been developed on the basis of probability theory, recent research has gone **beyond traditional approaches** and also **leverages more general formalisms and uncertainty calculi.**

8. BIAS & FAIRNESS

When **training** machine learning systems, the training **data can be biased**, leading to **unwanted outcomes**. For example, an HR system trained on old hospital personnel data might discriminate against women for doctor positions and against men for nurse positions, due to historical gender biases in these roles. This session will **explain** these issues, how to **avoid** them, how to **measure bias** and what the limitations of avoiding it are. Also advanced bias and fairness issues in large language models, and generative AI more generally, will be covered.

9. PRIVACY

Sometimes **the quality of machine learning system outputs and privacy are at odds and need to be balanced.** However, there are techniques that allow the training of machine learning systems on privacy sensitive data, without exposing the data itself. **Those techniques and relevant regulation on these practices are explained in this session.**

10. USE CASES

During the last session, some specific use cases in the domain of Explainable and Trustworthy AI will be discussed.

PRACTICAL INFORMATION

Fee

Onsite

The fee for onsite participation (in the UGain classroom) is **1.750 euro.** This includes the tuition fee, course notes, access to the digital e-learning environment, soft drinks, coffee and sandwiches.

Online

The for online participation only is **1.450 euro.**
This includes tuition fee and online access to the live sessions and the digital e-learning environment with digital course notes and recorded lessons.

Payment occurs after reception of the invoice.

All invoices are due in thirty days. All fees are exempt from VAT.

Reduction

When a participant of a company subscribes for the complete course, a reduction of 20% is given to all additional subscriptions from the same company. In that case, only one invoice is issued per company.

Cancellation policy

Our cancellation conditions can be consulted on
www.ugain.ugent.be/cancellation

Training vouchers

Ghent University accepts payments by KMO-portefeuille (www.kmo-portefeuille.be; authorisation ID: DV.0103194).

INFO AND SUBSCRIPTION

WWW.UGAIN.UGENT.BE/ EXPLAINABLEAI

GHENT UNIVERSITY

FACULTY OF ENGINEERING AND ARCHITECTURE

FACULTY OF BIOSCIENCE ENGINEERING

Time and location

- The lessons are given from 17h30 till 21h, with a sandwich break..
- Location: **Ghent University, UGain, building 60, Technologiepark Zwijnaarde.**
- The lessons can be followed onsite or online.
- Dates may change due to unforeseen reasons.

PROGRAMME

30 September 2024	1. INTRODUCTION Femke Ongenae & Sofie Van Hoecke
7 October 2024	2. WHITE BOX MODELS Daniel Peralta Cámara
14 October 2024	3. INTERPRETABILITY & EXPLAINABILITY Yvan Saeys
21 October 2024	4. ONLINE & TRANSFER LEARNING Matthias Feys
4 November 2024	5. HYBRID AI Femke Ongenae & Sofie Van Hoecke
18 November 2024	6. ROBUSTNESS Jonathan Peck
25 November 2024	7. UNCERTAINTY Willem Waegeman
2 December 2024	8. BIAS & FAIRNESS Tijl De Bie
9 December 2024	9. PRIVACY Tijl De Bie
16 December 2024	10. USE CASES Tijl De Bie, Matthias Feys, Femke Ongenae & Sofie Van Hoecke

Language

English is used in all presentations and documentation.

Organisation

Ghent University
UGain (UGent Academie voor Ingenieurs)
Technologiepark 60
9052 Zwijnaarde
09 264 55 82
ugain@ugent.be - www.ugain.ugent.be

With the support of VAIA