POST-ACADEMIC COURSE

EXPLAINABLE & TRUSTWORTHY ARTIFICIAL INTELLIGENCE

21 MARCH 2022 – 20 JUNE 2022

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.**

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 (limitation of 24 participants) or online (live and on demand) (unlimited).

CERTIFICATE

To receive a certificate, one should attend all the lessons and succeed for the final exam.
The exam will take place on September 12, 2022.

SCIENTIFIC COORDINATION

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

TEACHERS

- **prof. Tijl De Bie,** Department of Electronics and Information Systems, Ghent University
- **ir. Matthias Feys,** ML6
- **dr. Arne Gevaert,** Department of Applied Mathematics, Computer Science & Statistics, Ghent University
- **prof. Femke Ongenae,** Department of Information Technology, Ghent University
- **dr. Daniel Peralta Cámara,** Department of Information Technology, Ghent University
- **dr. Jonathan Peck,** Department of Applied Mathematics, Computer Science & Statistics, Ghent University
- **dr. Jan Van Looy,** ML6

INFO AND REGISTRATION

WWW.UGAIN.UGENT.BE/EXPLAINABLEAI

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.

Teachers: Femke Ongenae & Sofie Van Hoecke
Date: 21 March 2022

2. GLOBAL VS LOCAL INTERPRETABILITY

Machine learning systems help us connect cause and effect in complex data sets. **How we can make those interpretations depends on the type of algorithm used?** The level of interpretability we can reach will also differ depending on whether we look at how the model predicts in general, versus how a specific prediction of the model was computed. When interpretability is good enough, we can use hypothetical questions like "What if instead this would be the case?", adding to the usefulness of your system. **This lesson explains how to increase and measure those interpretabilities.**

Teacher: Arne Gevaert
Date: 28 March 2022

3. 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, decision Trees and Rule Sets and generalized Additive Models (GAMs).**

Teacher: Daniel Peralta Cámara
Date: 25 April 2022

4. SALIENCY MAPPING

To help explain how a neural network reached a certain conclusion, **certain visualizations of its reasoning can be useful.** One type of visualization is a heatmap showing which areas of a photo contribute most to how a system has labeled it. **These visualizations are explained in this session.**

Teacher: Arne Gevaert
Date: 2 May 2022

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.

Teachers: Femke Ongenae & Sofie Van Hoecke
Date: 9 May 2022

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**.

Teacher: Jonathan Peck
Date: 16 May 2022

7. 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.**

Teachers: Matthias Feys & Jan Van Looy
Date: 23 May 2022

8. BIAS & FAIRNESS

When training machine learning systems, **the training data can be biased, leading to unwanted outcomes,** i.e. an HR system trained on old hospital personnel data now stating that women might be unlikely good candidates for doctor positions. **This session will explain these issues, how to avoid them, how to measure bias and what the limitations of avoiding it are.**

Teacher: Tijl De Bie
Date: 30 May 2022

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.**

Teacher: Tijl De Bie
Date: 13 June 2022

10. USE CASES

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

Date: 20 June 2022

PRACTICAL INFORMATION

Fee

Onsite

The fee for onsite participation (in the UGain classroom) is **1.650 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 **1.400 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

Time and location

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

Language

English is used in all presentations, exercises 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

VAIA

GHENT UNIVERSITY

FACULTY OF ENGINEERING AND ARCHITECTURE

FACULTY OF BIOSCIENCE ENGINEERING