Teaching Guide					
Identifying Data					2022/23
Subject (*)	Explainable and Trustworthy AI			Code	614544004
Study programme	Máster Universitario en Intelixencia Artificial				
Descriptors					
Cycle	Period	Year		Type	Credits
Official Master's Degree	2nd four-month period	First		Obligatory	3
Language	English				
Teaching method	Face-to-face				
Prerequisites					
Department					
Coordinador	Alvarez Estevez, Diego		E-mail	diego.alvareze@udc.es	
Lecturers	Alvarez Estevez, Diego		E-mail	diego.alvareze@udc.es	
Web	www.usc.gal/gl/estudos/masteres/enxenaria-arquitectura/master-universitario-intelixencia-artificial/20222023/ia-explicable-confi				
General description	The main objective of this subject is to train students in the development of skills for an adequate treatment of privacy, reliability, transparency and interpretability of models and results associated with intelligent systems. Special emphasis will be placed on identifying and analyzing biases and their impact on the design of Artificial Intelligence (AI) algorithms. In addition to technical aspects, disruptive technologies and specific and general computer tools, aimed at covering all phases of the design, analysis and evaluation of intelligent systems, students will learn to know and understand the social and ethical implications of technology in general and of AI in particular Teaching guide coordinating center (USC): https://www.usc.gal/gl/estudos/masteres/enxenaria-arquitectura/master-universitario-intelixencia-artificial/20222023/ia-explicable-confiable-18828-17979-2-102310				

Study programme competences / results	
Code	Study programme competences / results
A6	CE05 - Ability to design and develop intelligent systems through the application of inference algorithms, knowledge representation and automated planning
A7	CE06 - Ability to recognise those problems that require a distributed architecture, not predetermined during the system design, suitable for the implementation of multiagent systems
A8	CE07 - Ability to understand the consequences of the development of an explainable and interpretable intelligent system
A9	CE08 - Ability to design and develop secure intelligent systems, in terms of integrity, confidentiality and robustness
B1	CG01 - Maintaining and extending theoretical foundations to allow the introduction and exploitation of new and advanced technologies in the field of AI
B2	CG02 - Successfully addressing each and every stage of an AI project
B3	CG03 - Searching and selecting that useful information required to solve complex problems, with a confident handling of bibliographical sources in the field
B6	CB01 - Acquiring and understanding knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, frequently in a research context
B7	CB02 - The students will be able to apply the acquired knowledge and to use their capacity of solving problems in new or poorly explored environments inside wider (or multidisciplinary) contexts related to their field of study
B8	CB03 - The students will be able to integrate different pieces of knowledge, to face the complexity of formulating opinions (from information that may be incomplete or limited) and to include considerations about social and ethical responsibilities linked to the application of their knowledge and opinions
B9	CB04 - The students will be able to communicate their conclusions, their premises and their ultimate justifications, both to specialised and non-specialised audiences, using a clear style language, free from ambiguities
C2	CT02 - Command in understanding and expression, both in oral and written forms, of a foreign language

C3	CT03 - Use of the basic tools of Information and Communications Technology (ICT) required for the student's professional practice and learning along her life
C4	CT04 - Acquiring a personal development for practicing a citizenship under observation of the democratic culture, the human rights and the gender perspective
C5	CT05 - Understanding the importance of the entrepreneurial culture and knowledge of the resources within the entrepreneur person's means
C6	CT06 - Acquiring abilities for life and healthy customs, routines and life styles
C7	CT07 - Developing the ability to work in interdisciplinary or cross-disciplinary teams to provide proposal that contribute to a sustainable environmental, economic, political and social development
C8	CT08 - Appreciating the importance of research, innovation and technological development in the socioeconomic and cultural progress of society

Learning outcomes			
Learning outcomes	Study programme competences / results		
Develop capacities for an adequate treatment of privacy, reliability, transparency and interpretability of models and results	AC5 AC6 AC7 AC8	BC1 BC2 BC3 BC6 BC7 BC8 BC9	CC2 CC3 CC4 CC5 CC6 CC7 CC8
Identify and analyze biases and their impact on the design of Artificial Intelligence algorithms	AC5 AC6 AC7 AC8	BC1 BC2 BC3 BC6 BC7 BC8 BC9	CC2 CC3 CC4 CC5 CC6 CC7 CC8
Know and understand the social and ethical implications of technology in general and Artificial Intelligence in particular	AC5 AC6 AC7 AC8	BC1 BC2 BC3 BC6 BC7 BC8 BC9	CC2 CC3 CC4 CC5 CC6 CC7 CC8

Contents	
Topic	Sub-topic
Explainability and interpretability. Model-agnostic methods. Explanations based on examples. FAT-E (fairness, accountability, transparency and ethics). Study and types of biases. Types and models of explanation. Evaluation methodologies. Data integrity, privacy, confidentiality and robustness of models. Reliability by design	Explainability and interpretability. Model-agnostic methods. Explanations based on examples. FAT-E (fairness, accountability, transparency and ethics). Study and types of biases. Types and models of explanation. Evaluation methodologies. Data integrity, privacy, confidentiality and robustness of models. Reliability by design

Planning

Methodologies / tests	Competencies / Results	Teaching hours (in-person & virtual)	Student?s personal work hours	Total hours
Laboratory practice	A6 A7 A8 A9 B1 B2 B3 B6 B7 B8 B9 C2 C3 C4 C5 C6 C7 C8	11	43	54
Guest lecture / keynote speech	A6 A7 A8 A9 B1 B2 B3 B6 B7 B8 B9 C2 C3 C4 C5 C6 C7 C8	10	10	20
Personalized attention		1	0	1
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.				

Methodologies	
Methodologies	Description
Laboratory practice	The interactive classes will take place in the selected Computer Classroom at each University, using various software tools for each of the thematic blocks, addressing exercises and projects with different levels of complexity. The students will work in individual positions with the constant support of the teaching staff. The scripts of the practices will be self-explanatory, allowing students to take profit of their personal work hours. Practical classes are aimed for developing skills CG1, CG3, CB6, CB7, CB8, CT3, CT8, CE5, CE6, CE7, CE8, CE9. Students develop practical work that involves dealing with the resolution of complex problems, and the analysis and design of solutions that constitute a means for their resolution. Students may have to present their work orally. The work done by the students can be done individually or in work groups. Students can work on the solution to the problems raised individually or in groups. This teaching methodology will be applied to the training activity "Practical laboratory classes" and may be also applied to the training activity of "Problem-based learning sessions, seminars, case studies and projects". Laboratory practices: the teaching staff of the subject poses to the students problems of a practical nature whose resolution requires the understanding and application of the theoretical-practical contents included in the contents of the subject. Learning by projects: students are presented with practical projects whose scope requires an important part of the total dedication of the student in this subject. In addition, due to the scope of the work to be carried out, it is required that the students apply technical and non-technical skills. Teaching will be supported by the virtual platform of the master in the following way: repository of documentation related to the subject (texts, presentations, etc.) and virtual tutoring of students (e-mail and forums). Tutoring: the teaching staff will assist students in individualized tutorial sessions dedicated to study orientation and the resolution of doubts about the contents and work of the subject

Guest lecture / keynote speech	The teaching methodology will be based on the individual work of the students, on the discussion with the teacher in class and on individual tutorials. Theory classes (expository): Oral presentation complemented with the use of audiovisual media and the introduction of some questions addressed to students, in order to transmit knowledge and facilitate learning. In addition to the oral presentation, students should dedicate some time to prepare and review the class materials on their own. For each theme or thematic module of the expository classes, the teaching staff will prepare the contents, explain the objectives of the theme to the students in class, present each theme with the aim of providing a set of information with a specific scope, suggest a bibliography, provide additional work material, etc. This teaching methodology will be applied to the training activity "Theory classes". Theory classes are aimed for developing skills CG1, CG3, CB6, CB7, CB8, CB9, CE5, CE6, CE7, CE8, CE9. In addition, the teaching staff will propose to the students a set of activities to carry out, individually or in groups (case studies, papers, presentations, readings, etc.). Students must submit a selection of them for evaluation. As a result, students will develop the skills CG3, CB7, CB8, CB9, CT2, CT3, CT4, CT6, CT8, CE7, CE8

Personalized attention	
Methodologies	**Description**
Laboratory practice	

Assessment			
Methodologies	**Competencies / Results**	**Description**	**Qualification**
Guest lecture / keynote speech	A6 A7 A8 A9 B1 B2 B3 B6 B7 B8 B9 C2 C3 C4 C5 C6 C7 C8	exam of the theoretical part (45%)	45
Laboratory practice	A6 A7 A8 A9 B1 B2 B3 B6 B7 B8 B9 C2 C3 C4 C5 C6 C7 C8	evaluation of the deliveries associated with the interactive sessions (35%), the delivery of a personal work and its oral presentation (15%) and the continuous assessment of each student throughout the course (5%)	55

Assessment comments

The evaluation of the learning considers an exam of the theoretical part (45%) and the evaluation of the deliveries associated with the interactive sessions (35%), the delivery of a personal work and its oral presentation (15%) and the continuous assessment of each student throughout the course (5%).

It is mandatory to pass all parts (exam, practices, work, continuous evaluation), considering the following criteria:

1. Exam (45%): The theoretical content of the subject will be evaluated in a single exam to be taken on the official date. The exam will consist of questions related to all the topics of the program. The exam will be specially oriented to evaluate the comprehension of the knowledge exposed in the theory classes. The exam grade will be the weighted average of the modules of the subject, which will only be calculated in the case of having a grade equal to or greater than 4 in each module.

2. Practices (35%): There will be mandatory deliveries associated with the interactive sessions related to each theoretical module. The solutions proposed by the students to the proposed practices will be evaluated. The evaluation of practices can be carried out through a correction by the teacher, or a defense of the solution provided by the student in the form of an oral presentation of the solution developed. (Applicable to the results of the training activities "Practical laboratory classes", "Problem-based learning, seminars, case studies and projects" and "Carrying out supervised work"). The average grade will only be calculated in the case of having a grade greater than or equal to 4/10 in all deliveries. In addition, it is mandatory face-to-face attendance of at least 60% of the interactive classes.

3. Work (15%): Students must submit and present a personal work according to the calendar established at the beginning of the semester. The evaluation of the supervised work will be carried out by means of a defense in which the students explain their proposal and conclusions to the teacher, or by means of an oral presentation of the solution in front of the classroom. The grade obtained will be the average of the evaluation of the written work and its oral presentation. The average will only be made if a grade equal to or greater than 4 is obtained in each part.

4. Continuous evaluation (5%): The attendance and active participation of students will be taken into account in the expository classes but also during the presentation of works, discussions, seminars, and in the interactive sessions that are held throughout the course. It is mandatory attending at least 60% of the presentation sessions and seminars.

The final grade for the subject will be the sum of the four partial grades, except in those situations indicated above. When any part is not passed, the final grade for the opportunity will be the minimum of the partial grades.

Students who have not participated in any of the evaluation activities will obtain the grade of not presented.

Students who have official exemption from class attendance must take, in any case, the final written exam, as well as doing all deliveries of practices and work that are established as mandatory throughout the course and, if required, make their oral presentations. In this modality, the tutoring, deliveries and oral presentations can be made remotely.

In the second opportunity, the students must pass the pending evaluation activities of the first opportunity, in accordance with the previous criteria.

For cases of fraudulent completion of exercises or tests, the provisions of the Regulations for evaluating the academic performance of students and reviewing grades will apply. The total or partial copy of any practice or theory exercise will automatically mean a grade of 0.0 in the subject and opportunity

Sources of information	
Basic	Supplementary material will be provided in the virtual platform of the master to facilitate following each unit in this subject. Given the heterogeneity of topics to be dealt within this subject, with each class, references to bibliographic resources as well as other types of content (reports, multimedia, etc.) will be provided to student for the more specific aspects of the subject. The following references are of a complementary type, they deal with general aspects related to explainable and trustworthy AI. 1. V. Dignum. Responsible Artificial Intelligence. How to Developand Use AI in a Responsible Way. Springer Nature Switzerland AG, 2019, ISBN: 978-3-030-30370-9 , https://doi.org/10.1007/978-3-030-30371-6 2. A. Barredo Arrieta et al., Explainable Artificial Intelligence(XAI): Concepts, taxonomies, opportunities and challenges toward responsibleAI, Information Fusion, 58:82-115, Elsevier 2020, https://doi.org/10.1016/j.inffus.2019.12.012 3. T. Miller, Explanation in artificial intelligence: Insights fromthe social sciences. Artificial Intelligence, 267:1-38, Elsevier 2019, https://doi.org/10.1016/j.artint.2018.07.007 4. G. Vilone, L. Longo, Notions of explainability and evaluationapproaches for Explainable Artificial Intelligence, Information Fusion,76:89-106, Elsevier 2021, https://doi.org/10.1016/j.inffus.2021.05.009 5. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti,D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACMComputing Surveys, 51(5):1?42, 2019, https://dl.acm.org/doi/10.1145/3236009 6. J.M. Alonso, C. Castiello, L. Magdalena, C.Mencar, Explainable Fuzzy Systems. Paving the way from interpretable fuzzysystems to explainable AI systems. SpringerInternational Publishing, 2021, ISBN: 978-3-030-71098-9, https://doi.org/10.1007/978-3-030-71098-9Supplementary material will be provided in the virtual platform of the master to facilitate following each unit in this subject. Given the heterogeneity of topics to be dealt within this subject, with each class, references to bibliographic resources as well as other types of content (reports, multimedia, etc.) will be provided to student for the more specific aspects of the subject. The following references are of a complementary type, they deal with general aspects related to explainable and trustworthy AI. 1. V. Dignum. Responsible Artificial Intelligence. How to Developand Use AI in a Responsible Way. Springer Nature Switzerland AG, 2019, ISBN: 978-3-030-30370-9 , https://doi.org/10.1007/978-3-030-30371-6 2. A. Barredo Arrieta et al., Explainable Artificial Intelligence(XAI): Concepts, taxonomies, opportunities and challenges toward responsibleAI, Information Fusion, 58:82-115, Elsevier 2020, https://doi.org/10.1016/j.inffus.2019.12.012 3. T. Miller, Explanation in artificial intelligence: Insights fromthe social sciences. Artificial Intelligence, 267:1-38, Elsevier 2019, https://doi.org/10.1016/j.artint.2018.07.007 4. G. Vilone, L. Longo, Notions of explainability and evaluationapproaches for Explainable Artificial Intelligence, Information Fusion,76:89-106, Elsevier 2021, https://doi.org/10.1016/j.inffus.2021.05.009 5. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti,D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACMComputing Surveys, 51(5):1?42, 2019, https://dl.acm.org/doi/10.1145/3236009 6. J.M. Alonso, C. Castiello, L. Magdalena, C.Mencar, Explainable Fuzzy Systems. Paving the way from interpretable fuzzysystems to explainable AI systems. SpringerInternational Publishing, 2021, ISBN: 978-3-030-71098-9, https://doi.org/10.1007/978-3-030-71098-9
Complementary	

Recommendations
Subjects that it is recommended to have taken before
Subjects that are recommended to be taken simultaneously
Subjects that continue the syllabus
Other comments

It
is recommended to bring the subject up to date and the use of tutoring sessions
to clarify doubts and get advise on its development. In addition, it is
recommended that students solve, verify and validate all the exercises and
practices proposed during the course (no matter if they are or not to be
officially evaluated)