5ARC0 Study Guide 2025 (Draft)

Class Schedule (See below for readings details)

Date/ time/ type	Topic, Activity, or Deadline	Readings (see list)
2-Sep/ H5-6/ Lecture	Ethics, human rights and AI	[UDHR]
2-Sep/ H7-8/ Tutorial	Multi-disciplinarity and group formation	
4-Sep/ H1-2/ Lecture	Philosophy of AI	[CHLEP Ch.1, Sections 1.1-1.3 & 1.10-1.13; Ch. 2]
8-Sep/ H9-10	ONLINE Client Meetings for topics EC (17:30) and AIER (18:00)	
9-Sep/ H5-6/ Lecture	Ethics of AI	[CHLEP Ch. 3]
9-Sep/ H7-8/ Tutorial	Literature search	
11-Sep/ H1-2/ Lecture	Fairness and AI	[CHLEP Ch. 4]
11-Sep/ H3-4	First Round of Student Meetings/ Client Meetings Client meeting for topic EDS (12:00)	
15-Sep/ H9-10	First Round of Student Meetings (ONLINE)	
16-Sep/ H5-6/ Lecture	Responsibility and Responsible AI	[CHLEP Ch. 5]
16-Sep/ H7-8/ Tutorial	First Round of Student Meetings	
18-Sep/ H1-2/ Lecture	Sustainability and AI	[CHLEP Ch. 6]
18-Sep/ H3-4	First Round of Student Meetings	
19-Sep 23:59	Assignment Phase 1 Due	
23-Sep/ H5-6/ Lecture	Explainable AI and the Black Box Problem	Zednik 2021
23-Sep/ H7-8/ Tutorial	Second Round of Student Meetings	
25-Sep/ H1-2/ Lecture	Human-Centered AI and Meaningful Human Control	Shneiderman 2020; Santoni de Sio & van den Hoven 2018
25-Sep/ H3-4	Second Round of Student Meetings	
29-Sep/ H9-10	Second Round of Student Meetings (Online)	
30-Sep/ H5-6/ Lecture	Second Round of Student Meetings (Online)	
30-Sep/ H7-8/ Tutorial	No Meeting at this time	
2-Oct/ H1-2/ Lecture	No Lecture/meeting.	
7-Oct/ H5-6/ Lecture	Machine Ethics and Autonomous Systems	Goodall 2014; Awad 2020

7-Oct/ H7-8/ Tutorial	Work time (Nickel)	
9-Oct/ H1-2/ Lecture	Superintelligence (Nickel)	Chalmers 2010 Sections 1, 2, 5, 6, 7
9-Oct/ H3-4	Third Round of Student Meetings with Practice Pitches - Atlas 10.301	
12 Oct 23:59	Submit Draft of Phase 2 for Peer Review	
13-Oct/ H9-10	Third Round of Student Meetings with Practice Pitches (Online)	
14-Oct/ H5-6/ Lecture	Client Final Presentation Session AIER (Online)	
14-Oct/ H7-8/ Tutorial	Final Client Presentation Session EDS-Alpha 0.98	
15-Oct/ H9-10 (Wednesday)	Third Round of Student Meetings with Practice Pitches (Online)	
16-Oct/ H3-4/ Lecture	Client Presentation Session EC (Online)	
17 Oct 23:59	Complete Peer Review (Individual)	
21-Oct/ H5-6	Office hour	
21-Oct 23:59	Submit Video Pitch	
23-Oct/ H1-2	Exam review	
24 Oct 23:59	Submit Phase 2 Final Version	
25 Oct 23:59	Complete Groupmate Evaluation	

[UDHR] United Nations, Universal Declaration of Human Rights. (1948). Available at https://www.un.org/en/about-us/universal-declaration-of-human-rights

[CHLEP] Smuha NA, ed. (2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge Law Handbooks. Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-the-law-ethics-and-policy-of-artificial-intelligence/contents/B58479AF2A9C398BD461575A18F828D3

Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, *63*(3), 48-55.

Chalmers, D.J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17: 7-65. Available at https://consc.net/papers/singularityjcs.pdf

Goodall, N. J. (2014). Ethical Decision Making During Automated Vehicle Crashes. Transportation Research Record: Journal of the Transportation Research Board, No. 2424, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 58–65. DOI: 10.3141/2424-0

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, *International Journal of Human–Computer Interaction* 36:6, 495-504, DOI:10.1080/10447318.2020.1741118

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable AI. *Philosophy & Technology* 34: 265-288.

There are also recommended readings for each of the project topics, which will be available on Canvas Pages.

Official course information from OSIRIS:

1. Basic course information

Academic year	2025-2026
Quartile	1, timeslot E
Course code	5ARC0
Course name	Human and ethical aspects of AI
Credits (ECTS)	5
Course category	not applicable
Course type	graduate school
Target group	master AI&ES students
Language of instruction	English
Research group	IE&ES groups: Philosophy and Ethics 100%
Responsible lecturer	dr. Philip Nickel
Contact person	dr. Philip Nickel
Co-lecturer(s)	TBA

2. Motivation

Core course in master program AI&ES

3. Aims / Learning goals

After following this course, the student is able to:

Anticipate and evaluate how AI affects human behavior, and vice versa.

Recognize and assess ethical problems in AI and develop a plan to find solutions in cooperation with stakeholders:

- Identify ethical issues and value conflicts at different stages of the design process.
- Describe these issues and conflicts as they relate to intelligent systems, using ethical frameworks and concepts.
- Adopt a strategy for resolving value conflicts in a way that is convincing to multidisciplinary stakeholders.
- Use these reflections to justify choices in the design of intelligent systems and interfaces.

4. Contents

Description of the course contents

This course addresses the human and ethical impacts of AI, giving students concepts and approaches to understand how AI and humans interact and to anticipate ethical issues that arise due to the application of AI to real-life problems.

AI affects human behavior and vice versa: AI can support humans in repetitive, easy to automatize tasks, reducing human error and boredom, as well as for complex tasks, extending their cognitive abilities, and improving or augmenting human decision making. However, AI can lead to job erosion, loss of skills, overreliance on automation, and other effects that should be prevented or mitigated.

Students learn to give reasoned answers to central ethical questions about AI such as: What is an ideal vision for the future of AI and automation? How can we assure that people affected by AI have meaningful control over artificial intelligence and robotics? How can professionals working in the field respond to ethical challenges that occur in their work? Topics include: ethical principles and theories, meaningful human control, moral responsibility, trust in AI, opacity and explainability, ethics of generative artificial intelligence.

ECTS breakdown: description of the ECTS distribution over the different content elements and how much time a student will spend on each content element

2.5 ECTS: students acquire theoretical understanding of ethics concepts and their application to AI.

Following lectures, reading and supervised tutorials, self-study, taking practice quizzes and exam

2.5 ECTS: students motivate an approach to a real-life problem case involving ethics and AI, and develop a strategy for addressing it with stakeholder input.

Supervised tutorials, student study group meetings, Internet research, writing

5. **Instruction methods (describe lab work separately)**

Lectures, supervised tutorials, self-study, practice quizzes, presentations
Student study groups focusing on the written assignment

6. **Materials**
   a. **Required materials.** None
   b. **Recommended materials** See Literature above

7. **Position in the curriculum**
   a. **Entrance requirements.** None
   b. **Assumed previous knowledge.** Basic ethics concepts
   c. **Assumed previous knowledge can be gained by.** 0SAB0 USE basic: ethics and history of technology or ITEC
   d. **Follow up courses**. 0HM340 Human-AI Interaction
   e. **Is the course part of a coherent package? If so, please state which coherent package.** No

8. **Assessment method (tests and weights)**

Written assignment (50%) & written (essay) exam (50%)

The written assignment is an interim assignment in which the students go into detail motivating an approach to a real-life problem, based on an analysis of scientific literature, experimentation with solutions, and consultation with domain experts.

There will be three possible topics, which will be further specified by each group:

a. AI-generated ethics cases [EC]
b. Artificial intelligence for ethics review [AIER]
c. AI for eating disorder symptoms [EDS]

The written assignment will be assessed primarily on awareness of ethical concerns raised by different approaches, motivation for iterative solutions to the problem, and sophistication in identification and anticipation of limitations and barriers in the target context. Clarity and compellingness of the design solution will also be assessed.

The written assignment includes Phase 1 which is a first version (10%). It also includes a peer review between groups, a presentation (10%) and an intra-group peer evaluation and reflection (10%).

In case the written assignment as a whole is evaluated with an insufficient, students have the opportunity to improve the main part of the written assignment based on the feedback. The improved paper has to be handed in no later than week 4 of Q2. The presentation and peer evaluation cannot be re-submitted.

Written (essay) exam means an on-campus exam, in which the students are expected to write out their answers (i.e., most answers will be open format). The exam tests competency in identifying and applying ethics and psychology concepts from the lectures. An example would be the ability to explain, apply, and critique a notion of trust, fairness, or manipulation as applied to AI systems. Students might be asked to reproduce central concepts from the lectures, or they might be given a case to which they must apply one or more concepts.

Students must obtain a grade of at least 5,0 on the written exam in order to pass the course.

9. **Other information.** Use of AI generated text and work in the course.

In any use of AI, you remain responsible for the work you deliver, down to the individual sentence.

Use of AI other than that indicated may lead to a suspicion of fraud.

Regarding the written assignment, you are allowed to use AI as part of a demonstration or mock-up. In other words, it is the object of your assignment. AI is not a means of explaining your process of working with that object or justifying your thought process. You may use it in superficial ways to brainstorm ideas or do a spelling and editing check, at your own risk. You should maintain complete critical control over its use.