Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Department of Philosophy

# ETHICS OF ARTIFICIAL INTELLIGENCE

## Curriculum (Syllabus)

### Course details

| | |
|---|---|
| Level of higher education | Second (educational and scientific), Master's degree |
| Field of knowledge | C - social sciences, journalism, information and international relations |
| Specialisation | C5 Sociology |
| Educational programme | Social Data Analytics |
| Status of discipline | Elective |
| Form of study | Full-time (day) |
| Year of study, semester | 1 year, 2nd semester |
| Number of ECTS credits | 150 hours (30 lectures, 30 practical classes, 90 independent study) |
| Semester assessment/assessment measures | Modular control work,<br>exam |
| Class schedule | https://schedule.kpi.ua/ |
| Language of instruction | Ukrainian / English |
| Information about course director / lecturers | Lecturer, Practical / Seminar: Kazakov Mstislav Andriyovych |
| Course location | Distance learning platform Sikorsky / Google classroom; link |

### Curriculum

#### 1. Description of the course, its purpose, subject matter and learning outcomes

Over the past 15 years, the greatest progress in big tech has been demonstrated by research and development in the field of machine learning, better known today as "artificial intelligence" — a set of techniques, technologies, methods for creating and deploying "smart" machines and programs — that is, those that partially reproduce one or another ability or active property of intelligence. Computer vision, image recognition and visual reasoning, natural language processing, and generative algorithms are the main areas of machine learning in which progress is most noticeable, serving as a prerequisite for the emergence of foundational models, known as GPAIs (General Purpose Artificial Intelligence Systems). Thanks to them, in particular through the deployment of large language models at the infrastructure and public levels, AI has not only become a mass phenomenon today, but has also been commercialised. Still infinitely far from the so-called "General Intelligence," synthetic entities have learned very well to emulate anthropomorphism itself — not only the external form, but also the "inner world" of a person: thoughts, the process of reasoning and conclusion, mental states (beliefs, knowledge, certainties or doubts), feelings, emotions; the most famous and sought-after area of emulation today is undoubtedly the products of intellectual activity, creativity in its broadest sense: from scientific research to musical compositions. The introduction of AI systems into everyday life, professional activities, logistics, manufacturing, administrative, financial and legal processes is inevitable and promises humanity economic benefits, improved living standards, freeing up time and freedom from routine tasks; in the longer and more imaginary term, AI technology promises us solutions to global problems and the creation of a new, non-biological form of intelligent life, "sentient beings" known today as AGI – Artificial General Intelligence.

However, it is also a fact that these scalar ambitions and dreams are undergoing a scalar *collapse*, and the promised golden age is always postponed to an indefinite "tomorrow." In addition to the lack of technical (engineering) solutions and technological "lag" (the slow and uneven spread of automation by AI systems), and unlike other technologies with transformative potential (such as bitcoin), AI is the most problematic technology from *an ethical* point of view. The ethical conflicts of artificial intelligence cover a wide range of issues – from existential (speculative) risks, loss of cognitive autonomy and unemployment to the centralisation of goods and resources, increased inequality, the entrenchment of existing power structures, systems of disciplinary practices, forced "normalisation," automation of weaponisation, control, and surveillance of the population, to the use of algorithms as final agents in decision-making that directly affect human interests, freedom, autonomy, health, and life. In addition to humans, AI at its current stage of historical development also causes serious damage to the environment, leads to excessive resource consumption, and has a deep carbon footprint: training a single large language model has long since surpassed the aviation industry in terms of $CO_2$ emissions. AI systems are not a speculative threat to the future, but a series of structural and systemic problems of varying scales that need to be solved today.

The aim of studying this discipline is to demystify and demythologise artificial intelligence by providing the critical tools and critical-theoretical framework, methodology, technical-technological and ethical knowledge necessary for this, forming a balanced and synoptic view, attitude and skills necessary for interacting with it. Skills of this kind are necessary for both users and other stakeholder groups, as we are all sources of ethical problems: academic researchers, developers, owners, infrastructure providers, investors, interest groups (from marginalised communities to "average" consumers). National states, corporations, and the military are pursuing their own interests here, turning AI into an instrument of geopolitical influence and provoking an arms race, the result of which, in combination with the neoliberal "moment" of the present, has led to the current state of affairs we have in AI, a kind of "scalar Darwinism," a race among developers to constantly increase their transformer models without goals or infrastructural changes.

Starting with a fundamental deconstruction of the meaning of the concepts of "artificial" and "intelligence" in the humanistic, technological and scientific spectrums, while also indirectly reviewing some aspects and problems of computer science and development, philosophy, theology, history, literature, and applied industry and manufacturing, we will discover that "ethical AI" has never existed, dispelling the popular myth of "dreaming scientists who decided to teach machines to think like humans": from cybernetics and Alan Turing to the present day, artificial intelligence has been and remains the result of synergy at the intersection of interests, legislative and material capabilities, and ambitions of academia, the military, special services, national governments, and capitalist agents, from venture capital funds to megacorporations, the leviathans of so-called big tech. But if ethical AI did not exist, we must create it! Responsible and human-centred (not anthropocentric) AI is a priority both for us and for future generations of humans and non-humans, to whom we also have a responsibility. We are the main problem of the modern ethical challenges generated by this technology. But we are the ones who are capable of responding to these challenges, solving problems at the level of governance, design and development, intellectual and consumer decisions, and creating an ecosystem of sustainable and fair AI. An additional goal is to cultivate flexible and adaptive thinking in course participants when considering various problematic aspects of computer architecture and ways of implementing AI, perceiving the logical basis for AI decision-making, incorporating the knowledge and skills gained into a framework of social, civic and ethical responsibility through a range of options for actions, deeds, decisions, approaches, tactics and strategies that must be navigated in the current AI landscape in order to change it for the better.

After studying the course, students will be able to develop the following programme
learning outcomes:

a broad and unbiased understanding of the real state of affairs in the modern AI industry with knowledge of the history and genealogy – the origins of modern AI systems;

knowledge of the real history of AI – on the one hand, as a subject of science fiction and philosophical reflections and stories, and on the other, as the "fruit" of the military-industrial complex in collaboration with scientists.

Acquiring abilities and practical skills for the pragmatic implications of AI in such critically important areas and segments of human reality today as automation, law and legislation, military action, logistics, transport, industry, education, art, politics and AI management in general;

the ability to generate new knowledge, ideas, models, and hypotheses regarding long-term and highly abstract moral issues and dilemmas;
the ability to think speculatively and implicitly as forms of anticipating future ethical dilemmas;
include AI implications for human perception and personality, at the level of everyday and non-trivial computational intuitions regarding the operation of AI, both by specialists and those who are not specialists in computer science;
improving competence and orientation in literature, moral philosophy, management, and history;
the ability to conduct and organise theoretical debates, plenary groups and working groups to address practical, current issues related to various implementations of "narrow AI" in accordance with the real challenges posed by its exploitation in the social and individual spheres;
broad awareness, knowledge, skills and abilities to participate in speculative discussions and research on future implementations of AI and moving beyond "narrow" AI to General AI and Superintelligence – both as an existential risk and as a potential solution to all or part of the global problems of humanity as a whole;
critical thinking in conditions of radical uncertainty or absence of critically important pieces of information caused by the unpredictability – temporal and qualitative – of the development of artificial intelligence technologies and the emergent nature of their practical application.

## 2. Prerequisites and post-requisites of the discipline (place in the structural-logical scheme of training under the relevant educational programme)

*Prerequisites*: understanding of basic philosophical concepts and theories in the field of ethics, epistemology and metaphysics; understanding of the basics of computer science and research in the field of AI (machine learning, data structures, neural networks, applied AI, etc.) is desirable but not mandatory; understanding of key concepts of social and political theories, such as justice, public policy, governance, management, etc., will also be an advantage. Knowledge of English is another advantage, as a large amount of research and the latest news on AI ethics and industry development in general is mainly available in English.
*Post-requisites*: specific ethical issues (privacy, bias, agency, autonomy, AI technology management, regulation, etc.); AI policy and management, international and national management frameworks; research methods in AI, data analysis, quantitative and qualitative research, interdisciplinary and transdisciplinary research.

## 3. Course content

**Topic 1.** Artificial intelligence: history and forms of existence. Ethics: the philosophical science of moral systems, actions and duty. History and genealogy of AI ethics.
**Topic 2.** Theoretical basis of ethical problems of AI technologies from the point of view of computer science.
**Topic 3.** Scalar Darwinism as the logic of artificial intelligence development in the era of capitalist realism.
**Topic 4.** Moral and epistemic status of artificial agents. The problem of the responsibility gap.
**Topic 5.** Atlas of risks and threats posed by AI.
**Topic 6.** Privacy and ethics in data science and big data.
**Topic 7.** Countering threats and approaches to solving ethical problems.
**Topic 8.** Paradigms of ethical AI: from public to explainable.
**Topic 9.** Artificial intelligence as a source of global asymmetries and inequalities: social, educational, economic, epistemic, cognitive, workforce skills in labour markets
**Topic 10.** Principles of safe AI system design
**Topic 11.** The problem of value formation and implementation (alignment problem).
**Topic 12.** AI and the problem of control.
**Topic 13.** Artificial intelligence, ecology, long-termism (the "Green" AI paradigm).

## 4. Teaching materials and resources

Recommended teaching materials and resources are provided to help students master the material covered in lectures and practical classes. In accordance with the recommendations for creating "flexible educational architectures" with the aim of achieving didactic and accessible educational goals by including mixed and digital materials that can facilitate the process of mastering the material by non-specialist

students in the field of ICT (information and computer technologies), a list of recommended video materials (films, series, lectures, documentaries, YouTube content, materials from the Internet archive and videos with a Creative Commons licence, recordings of webinars and/or conferences), computer games, AI chatbots and image boards relevant to the course topic.

## Literature

### Basic


1. Avdeeva T. (2024) Actions and Dreams: Artificial Intelligence in the Public Sector - NGO "Digital Security Laboratory". – https://dslua.org/wp-content/uploads/2024/02/Mrii_ta_dii_kopiia.pdf

2. Bostrom, N. (2020) Superintelligence: Pathways, Dangers, and Ways to Prosper. Kyiv: Nash Format

3. Davenport, T., Corby, J. (2018) Job Vacancy: Human. How not to lose your job in the age of artificial intelligence. Kyiv: Nash Format

4. Kogut Yu.I. (2024) Artificial Intelligence and Security: A Practical Guide. Kyiv: Consulting Company "SIDCON"; VD Dakor

5. Kotsovsky, V.M. (2016) Methods and Systems of Artificial Intelligence. Lecture Notes. – https://shorturl.at/E16cf

6. Li, K.-F. (2020) Artificial Intelligence Superpowers: China, Silicon Valley and the New World Order. Kyiv: Force Ukraine.

7. Leach, N. (2024). Architecture in the Age of Artificial Intelligence. An Introduction to AI for Architects. ArtHuss

8. Methods and Systems of Artificial Intelligence: Textbook / Compiled by D.V. Lubko, S.V. Sharov. – Melitopol: FOP Odnorog T.V., 2019. – http://elar.tsatu.edu.ua/bitstream/123456789/15462/1/5_lubko_metody_2019.pdf

9. Russell, S. (2020) Human-Compatible: Artificial Intelligence and the Problem of Control. Kyiv: Force Ukraine

10. Seita de, M. (2023) The Code of Creativity. How artificial intelligence learns to write, draw, and think. Kyiv: Arthuss

11. Walsh, T. (2025). Fake: Artificial Intelligence in the Human World. Kharkiv: Fabula Publishing House

12. Sharov, S. (2023) The current state of artificial intelligence development and areas of application. - https://www.researchgate.net/publication/370595289_Sucasnij_stan_rozvitku_stucnogo_intelektu_ta_napramki_jogo_vikoristanna


### Supplementary


1. Almeida, I. (2024) Responsible AI in the Age of Generative Models. Ethics, Governance & Risk Management. Now Next Later AI.
2. Bednarz, Z., Zalnieriute, M. (Eds.) (2024) Money, Power, and AI. Automated Banks and Automated States. Cambridge University Press.www.doi.org/10.1017/9781009334297
3. Boddington, P. (n.d.). AI Ethics: A Textbook. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9382-4
4. Chetouani, M., Dignum, V., Lukowicz, P., & Sierra, C. (Eds.). (2023). Human-Centered Artificial Intelligence: Advanced Lectures. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24349-3
5. Coeckelbergh, M. (2020). AI Ethics. The MIT Press.
6. Crawford, K. (2021) Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence. Yale University Press.

7. Montemayor, C. (2023). The Prospect of a Humanitarian Artificial Intelligence: Agency and Value Alignment. Bloomsbury Academic Publishing.
8. O'Neil, C. (2016) Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy. Crown New York
9. Smuha, N.A. (Ed.) (2025) The Cambridge Handbook of The Law, Ethics and Policy of Artificial Intelligence. Cambridge University Press. https://www.sipotra.it/wp-content/uploads/2025/03/The-Cambridge-Handbook-of-the-Law-Ethics-and-Policy-of-Artificial-Intelligence.pdf

Literature from the list that may not be freely available for download at the time of request will be provided by the lecturer. Some of the seminar assignments are based on materials translated from English and adapted to the educational process.

**Additional content (video materials)**

1. 2001: A Space Odyssey. Directed by Stanley Kubrick, MGM, 1968.
2. Blade Runner. Directed by Ridley Scott, Warner Bros., 1982.
3. Blade Runner 2049. Directed by Denis Villeneuve, Warner Bros., 2017.
4. Ex Machina. Directed by Alex Garland, Universal Pictures, 2015.
5. Her. Directed by Spike Jonze, Warner Bros., 2013.
6. Black Mirror: "White Christmas." Created by Charlie Brooker, season 2, episode 4, Channel 4, 2014.
7. Black Mirror: "Hated in the Nation." Created by Charlie Brooker, season 3, episode 6, Netflix, 2016.
8. Black Mirror: "Be Right Back." Created by Charlie Brooker, season 2, episode 1, Channel 4, 2013.
9. AlphaGo. Directed by Greg Kohs, Little Men, 2017.
10. Coded Bias. Directed by Shalini Kantayya, 7th Empire Media, 2020.
11. The Social Dilemma. Directed by Jeff Orlowski, Netflix, 2020.
12. Humans Need Not Apply. CGP Grey, YouTube, 2014,www.youtube.com/watch?v=7Pq-S557XQU
13. Slaughterbots. The Future of Life Institute, YouTube, 2017,www.youtube.com/watch?v=9CO6M2HsoIA
14. Lo and Behold: Reveries of the Connected World. Directed by Werner Herzog, Magnolia Pictures, 2016.
15. Person of Interest: "If-Then-Else." Directed by Chris Fisher, season 4, episode 11, CBS, 2015.
16. The Cleaners. Directed by Hans Block and Moritz Riesewieck, Gebrueder Beetz Filmproduktion, 2018.
17. iHuman. Directed by Tonje Hessen Schei, UpNorth Film, 2019.
18. AI: Artificial Intelligence. Directed by Steven Spielberg, Warner Bros., 2001.
19. Ghost in the Shell. Directed by Mamoru Oshii, Production I.G, 1995.
20. You Look Like a Thing and I Love You: AI Weirdness Explains AI. Janelle Shane, YouTube, 2019,www.youtube.com/watch?v=H7H6Q7vPIGo

**Additional content**
**(interactive materials, repositories, forums, computer games)**

1. Absurd Trolley Problems. Neal Fun, 2019, https://neal.fun/trolleyproblems/
2. Can We Talk about AI Ethics? r/ArtificialSentience, Reddit, 2022, https://www.reddit.com/r/ArtificialSentience/comments/jkl345/can_we_talk_about_ai_ethics/
3. Cardboard Crash. Directed by Vincent McCurley and National Film Board of Canada, NFB, 2015, https://www.nfb.ca/film/cardboard_crash/

4. Debate: Is It Ethical to Work at AI Capabilities Companies? LessWrong, Aug. 2024, https://www.lesswrong.com/posts/abc456/debate-is-it-ethical-to-work-at-ai-capabilities-companies
5. Detroit: Become Human. Developed by Quantic Dream, Sony Interactive Entertainment, 2018, https://www.playstation.com/en-us/games/detroit-become-human/
6. Developing AI Safety: Bridging the Power-Ethics Gap. LessWrong, Apr. 2025, https://www.lesswrong.com/posts/def789/developing-ai-safety-bridging-the-power-ethics-gap
7. Eliezer Yudkowsky (Eliezer_Yudkowsky). " Pausing AI Developments Isn't Enough. We Need to Shut it All Down." LessWrong, 8 April 2023, https://www.lesswrong.com/posts/oM9pEezyCb4dCsuKq/pausing-ai-developments-isn-t-enough-we-need-to-shut-it-all-1
8. Ethics & Society Spaces. Hugging Face Spaces, 2023, https://huggingface.co/spaces/society-ethics
9. Ethics and Bias in AI. Hugging Face Spaces, 2024, https://huggingface.co/spaces/ethics-bias
10. Faceminer. Wristwork, 27 Feb. 2025, https://faceminer.com/
11. GitHub Topics: AI-Ethics. GitHub, 2025, https://github.com/topics/ai-ethics
12. IBMDeveloperMEA/AI-Ethics. GitHub repository, IBMDeveloperMEA, 2025, https://github.com/IBMDeveloperMEA/AI-Ethics
13. Identifying Bias in AI. Kaggle, 2023, https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai
14. In Your Experience, Are AI Ethics Teams Valuable/Effective? r/MachineLearning, Reddit, 2023, https://www.reddit.com/r/MachineLearning/comments/11sfhzx/in_your_experience_are_ai_ethics_teams/
15. MIT Moral Machine. Massachusetts Institute of Technology, 2016, https://moralmachine.mit.edu/
16. r/AIethics. Reddit, 2025, https://www.reddit.com/r/AIethics/
17. r/ArtificialSentience. Reddit, 2022, https://www.reddit.com/r/ArtificialSentience/
18. r/MachineLearning. Reddit, 2023, https://www.reddit.com/r/MachineLearning/
19. This Person Does Not Exist. https://thispersondoesnotexist.com
20. The Worst Internet-Research Ethics Violation I Have Ever Seen. The Atlantic, 2 May 2025, https://www.theatlantic.com/technology/archive/2025/05/reddit-ai-persuasion-experiment-ethics/682676/
21. Trolley Problem VR. Gaoqi, 2021, https://gaoqi.com/trolley-vr
22. Trolley Problem, Inc. Developed by Read Graves, Yogscast Games, 2022, https://store.steampowered.com/app/123456/Trolley_Problem_Inc/

## Educational content

### 5. Methodology for mastering the academic discipline (educational component)

The academic discipline covers 30 hours of lectures and 30 hours of practical classes, as well as the completion of a modular control work (MCW). Practical classes in the discipline are conducted with the aim of consolidating the theoretical provisions of the academic discipline and enabling students to acquire the skills and experience to operate with modern concepts in the field of artificial intelligence ethics. Based on the time allocated for studying the discipline, fourteen practical classes are recommended (taking into account the time for the MCT).

Based on the principles of flexible educational architecture, teaching methods and forms include not only traditional university lectures and seminars, but also elements of teamwork and group discussions. Active learning strategies are used, which are determined by the following methods and technologies: problem-based learning methods (research method); personality-oriented technologies, visualisation and information and communication technologies, in particular electronic presentations for lectures. Communication with the teacher is built using the "Electronic Campus" information system, the "Sikorsky" distance learning platform, as well as communication tools such as email and Telegram.

The ethical, social, and cultural implications of artificial intelligence are not limited to books, articles, and media news content, but are evident across the entire spectrum of contemporary digital culture: in games, interactive simulations, online forums, and video essays. With this in mind, the course offers a number of non-traditional resources as additional sources of theoretical and practical information, including films, computer games, interactive tools, forum discussions and subreddits, and digital artefacts that can function as full-fledged sources for analysing the phenomenological and mediated experience of AI interaction or interpretations.

## Lectures

Lecture 1. Artificial intelligence as an ethical problem

1. Ethics: philosophical study of morality, actions, and ways of life.
2. Artificial intelligence: a general perspective.
3. "Let's start with the fact that we have never been ethical" – AI as a "myth of givenness".
4. AI ethics, applied and speculative.
5. From "Erewhon" and "Dune" to "Westworld" and "Person of Interest": speculative science fiction as a source of the genesis of AI ethics.

Lecture 2. "Scalar Darwinism" as the logic of AI technology development

1. Large language models and accelerationism
2. Surveillance capitalism
3. AI capitalism
4. Capitalist realism, the neoliberal "moment," AI technologies
5. Scalar Darwinism as the logic of development and form of existence of modern artificial intelligence

Lecture 3. Theoretical foundations of ethical challenges of AI

1. Basic concepts and subject areas of artificial intelligence ethics
2. Implications of computer science: the ethical dimension
2.1. The theorem of no free lunches in search and optimisation and algorithmic bias
2.2. The theorem on the impossibility of machine fairness: it is impossible to agree on all criteria and take all wishes into account

Lecture 4. Artificial intelligences: real and possible
1. Levels of intelligence from narrow AI and microbial colonies to superintelligences
2. Functional implementations of AI systems
3. Typology of intelligent agents as an ethical problem
4. Ethical consequences of substantial implementations (tools, genies, oracles, sovereigns)

Lecture 5. Artificial moral agents and the gap in responsibility.

1. Can a machine be a moral agent?
2. Four "gaps": the problem of responsibility
3. The problem of the "Principal-Agent" (First and Second Orders).
4. Consequentialism, deontology, and metaethics on AI agenticity and agency

Lecture 6. Machine ethics in the era of large language models

1. Are LLMs agents: "Stochastic parrots" or "Candidates for sentience"?
2. Deflationism and inflationism, from reductionism to all-inclusive epistemology.
3. Emergent properties, explainability of "black boxes" as a field of contemporary speculative philosophical debate
4. Empirical turn in AI philosophy: research on interpretability as an evidence base for new arguments
5. Computational inferentialism: VMMs – agents without consciousness in the space of causes

Lecture 7. Risks and challenges of artificial intelligence.

1. Classifying the risks of artificial intelligence
2. General typology.

2.1. Discrimination and toxic content
2.2. Privacy and security
2.3. Misleading information
2.4. AI-Darkside
2.5. Risks of interaction
2.6. Society, economy, politics, ecology.
2.7. Security, system failures, internal limitations of AI systems.

Lecture 8. Privacy and fairness

1. Ethics for data science
2. Weapons of Math Destruction
3. Datasets and benchmarking: from narrow-minded people to "narrow-mindedness by design"
4. AI "superpowers" and global inequality

Lecture 9. Regulatory Framework for Artificial Intelligence in the World and State Strategies for AI Governance
1.1. EU
1.2. USA
1.3. United Kingdom
1.4. China
1.5. Southeast Asia
1.6. Africa
1.7. UN, G7 and other supranational actors
1.8. AI regulation in Ukraine

Lecture 10. Methods of countering ethical risks and challenges
1. Institutional methods
2. Technical methods
3. Transdisciplinary methods

Lecture 11. Making AI Ethical
1. Fairness (AI Fairness)
2. Transparency
3. Traceability
4. Accountability
5. Robustness
6. Resistance to attacks
7. Trustworthiness and alignment
8. Sustainable development
9. Freedom and autonomy

Lecture 12. Public AI: Prospects, goals, obstacles to the alternative to proprietary AI

1. Public AI ecosystem
2. General goals of public AI projects
3. Examples of goal implementation: Case studies
4. Directions for the development of public alternatives
5. Roles and recommendations: What stakeholders can *really* do
6. Tools: from standardisation to regulation

Lecture 13. Principles of safe AI system design

1. Design as a meta-ethical framework.

2. Principles of safe design and their implementation:
2.1. Redundancy
2.2. Transparency
2.3. Separation of duties
2.4. Minimisation of privileges
2.5. Fault tolerance
2.6. Anti-fragility
2.7. Negative feedback
2.8. Depth defence
2.9. Assignability
2.10. Discontinuity
2.11. Agential discernibility
2.12. Methods of controlling abilities
2.13. Domestication
3. Three synergies: Principles in action.

Lecture 14. The Coordination Problem: Values in AI Systems and Ways of Implementation

1. The problem of implementing values.
2. Metrics: Usefulness, Honesty, Harmlessness
3. Methods of implementing values
3.1. Prescriptive programming and its criticism
3.2. Reinforcement learning: from humans to models
3.3. Two ways of interactive learning of values
3.4. New ways of coordination (evolution, augmentation, self-game)
3.5. Moral parliament
4. Alternative frameworks: prospects for the future

Lecture 15. AI in the context of environmental ethics

1. Environmental ethics and the long-termism dilemma
2. Technology as a mediator of human-nature relations. AI as technology.
3. AI as a factor in environmental degradation
3.1. Energy consumption
3.2. Carbon footprint
3.3. Water resources
3.4. Resource extraction and hardware production
3.5. Non-environmentally friendly logistics
3.6. Internet of Things footprint
3.7. E-Waste
3.8. Indirect impacts
4. Our responsibility in the Anthropocene era
5. Towards Green (Sustainable) AI

## Practical (seminar) classes

The main objective of the seminar series is to deepen the knowledge students gain in lectures, develop their skills in working with basic and additional literature, form the ability to argue their own opinions, and develop communication skills. Seminar classes should contribute to a better understanding of the theoretical material from the course "Ethics of Artificial Intelligence". In order to integrate the theoretical component (knowledge) with the practical component (application of knowledge), the seminar questions are: 1) theoretical questions relevant to *the* lecture *material* that preceded the seminar classes; 2) independent case studies on the *topic* of the lecture material it covers, or based on *a sample* of this type of study presented during the relevant lecture *or* in the form of a sample template directly attached to the assignment conditions; 3) cases on the topics of the lectures, on which the applicants prepare and present

reports during the class, in which they present both their vision of an ethical problem in the field of AI and the solution they consider to be the best possible. For some selected topics, teamwork in groups of 2-3 people is provided. The distribution is as follows: 7 seminars are devoted to theoretical issues covered in lectures throughout the year, 8 seminars are devoted to case presentations and other forms of practical activity.

**Seminar 1.** Genealogy and history of artificial intelligence.
**Seminar 2.** Computer science theorems and machine ethics
**Seminar 3.** "Scalar Darwinism" of modern AI systems and questions of alternatives
**Seminar 4.** AI, the four gaps of responsibility and the two problems of the "agent-principal"
**Seminar 5.** Modern large language models as a subject of ethical research
**Seminar 6.** Risks and threats of intelligent agents: *from* the future *to* the present
**Seminar 7.** Regulatory, managerial and legalistic approaches to solving ethical problems
**Seminar 8.** Engineering solutions and implementation of desired qualities
**Seminar 9.** Public AI: The commons as an alternative
**Seminar 10.** Risk modelling: AI ethics through the prism of risk management
**Seminar 11.** Principles of safe design: concept, implementation, synergy
**Seminar 12.** Problems of Control, Coordination, and Implementation of Values
**Seminar 13.** Privacy and ethical conflicts in data science
**Seminar 14.** AI and Environmental Ethics in the Anthropocene. Simulation of "Decisive Debates": A stakeholder summit that shapes all aspects of the AI agenda for the next 20 years, from regional and international legislation to the powers of auditors of private proprietary technologies and research areas that will be/cease to be funded.
**Seminar 15.** Modular control work

## 6. Independent work of higher education applicants

Independent work by higher education applicants as part of the educational component includes a set of thematic questions for reflection and practical tasks aimed at self-assessment of knowledge and organisation of self-preparation by applicants within each of the practical/seminar classes and involves: preparation for classroom classes with a theoretical component; preparation of research and presentation of case studies (practical component); preparation for modular control work; preparation for the exam; in-depth preparation for classes – studying additional literature and alternative primary sources listed in additional lists (optional); preparation for group work (optional).

| No. | Independent work by students | Number of hours |
|---|---|---|
| 1 | Preparation for classroom activities | 56 |
| 2 | Preparation for modular control work | 4 |
| 3 | Preparation for the exam | 30 |
| | Total | 90 |

# Policy and control

## 7. Academic discipline policy (educational component)

**Class attendance rules**

Attendance at lectures is compulsory. Higher education students who miss lectures may have difficulty preparing adequately for seminars, but they are not required to make up for missed lectures. During lectures, higher education students are advised to take notes on the main aspects, key concepts, definitions, classifications and algorithms explained by the lecturer.

Active participation in seminars (practical classes) is mandatory and is important for the formation of the student's rating. When preparing for a seminar, a higher education student must study the lecture

material on a specific topic, as well as the information on the topic presented in the main list of literature and problem cases that follow the lecture and are brought to the relevant practical classes as topics for presentation of answers. If there are any questions or unclear points, they should be discussed with the teacher. If a higher education applicant has not had time to prepare, they should listen carefully to the presentations of others and try to compensate for their lack of preparation by assimilating new information.

If a student misses seminars or tests for valid reasons (e.g., due to illness or other important circumstances), they will be able to complete the assignment during the following week. The knowledge of higher education students on the topics they have missed will be tested through consultations with the lecturer, the schedule of which is available on the website of the Department of Philosophy.

No points will be awarded for actual attendance at lectures and seminars.

**Distance learning**

In the event of the introduction of a distance (blended) learning format, the educational process will be organised in accordance with the Regulations on Distance Learning at Igor Sikorsky KPI (https://osvita.kpi.ua/index.php/node/188 ), Regulations for conducting semester control in a distance mode (https://osvita.kpi.ua/node/148).

The educational process is organised using distance learning technologies, in particular through the Sikorsky Distance Learning Platform (https://www.sikorsky-distance.org) and the Electronic Campus AS (https://ecampus.kpi.ua). Higher education students join the Sikorsky platform (Google Classroom) via corporate email in the @lll.kpi.ua domain.

The distance learning process is carried out in accordance with the approved schedule of classes. In distance learning mode, classes take place in the form of online conferences on the Zoom platform. A link to the conference is provided at the beginning of the semester.

Distance learning classes are conducted via online conferences on the Zoom platform. Assessment results are posted on the Electronic Campus AS on the personal page of the higher education applicant (https://ecampus.kpi.ua).

**Rules of conduct in class**

During classes, students must adhere to the standards of ethical behaviour set out in the Code of Honour of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (https://kpi.ua/code), as well as the Regulations on the Commission on Ethics and Academic Integrity of NTUU "KPI" (https://data.kpi.ua/sites/default/files/files/2015_1-140a1.pdf).

On the university premises, higher education students are required to comply with the established Internal Regulations (https://kpi.ua/admin-rule). During lectures and practical classes in classrooms, as well as during video conferences, mobile phones should only be used in silent mode and exclusively for educational purposes — to search for necessary information on the Internet.

**Rules for awarding incentive and penalty points**

Incentive points are not included in the main RSO scale.

The maximum number of incentive points is **10** (10% of the total rating points).

Higher education applicants can receive incentive (additional) points for participating in international and/or all-Ukrainian scientific, scientific-practical conferences, Olympiads (student, all-Ukrainian), and competitions on the subject of the academic discipline.

Penalty points are not provided for.

**Assessment policy for control measures**

Assessment of control measures is carried out in accordance with the Regulations on the system of assessment of learning outcomes at Igor Sikorsky KPI (https://osvita.kpi.ua/node/37), Regulations on current, calendar and semester assessment of learning outcomes at Igor Sikorsky Kyiv Polytechnic Institute (https://osvita.kpi.ua/node/32).

The results of semester assessment are posted on the AS "Electronic Campus" on the personal page of the higher education applicant (https://ecampus.kpi.ua).

If a higher education applicant disagrees with the assessment of the control measure results, they have the right to appeal on the day the results of the relevant assessment are announced to the dean of the

faculty in accordance with the procedure set out in the Regulations on Appeals at Igor Sikorsky Kyiv Polytechnic Institute (https://osvita.kpi.ua/index.php/node/182).

**Deadline and resit policy**

Failure to complete assignments or meet deadlines for invalid reasons will result in the loss of the opportunity to receive the corresponding rating points. In case of absence from control measures for valid reasons, the higher education applicant is given the right to complete the assignment within the next week.

The procedure for eliminating academic debt and retaking semester tests is regulated by the Regulations on current, calendar and semester testing of learning outcomes at Igor Sikorsky Kyiv Polytechnic Institute (https://osvita.kpi.ua/index.php/node/32). A higher education student who has academic debt as a result of semester exams also has the right to eliminate it in accordance with the Regulations on the provision of additional educational services to higher education students at Igor Sikorsky KPI (https://osvita.kpi.ua/index.php/node/177).

**Recognition of learning outcomes acquired in non-formal/informal education**

The procedure for recognising such results is regulated by the Regulations on the recognition of learning outcomes acquired in non-formal/informal education (https://osvita.kpi.ua/index.php/node/179).

Individual content modules or topics of a discipline may be credited. In this case, the applicant is exempt from completing the relevant tasks, receiving the maximum score for them in accordance with the rating assessment system.

**Extracurricular activities and involvement of professional practitioners**

During the study of an academic discipline, extracurricular activities are possible, including attendance at scientific and practical events, lectures, training sessions, etc., within the scope of the discipline.

**Academic integrity**

In the process of studying the academic discipline, it is necessary to strictly adhere to the academic integrity policy defined by current legislation and internal documents of the educational institution.

The policies, standards, and procedures for academic integrity are contained in the following regulatory documents of Igor Sikorsky KPI, which are published on the university's website: Code of Honour of Igor Sikorsky KPI (https://kpi.ua/files/honorcode.pdf), Regulations on the system for preventing academic plagiarism (https://rb.gy/agihij), regulatory and legal documents, official recommendations, orders and directives, sociological studies, methodological materials, educational courses (https://kpi.ua/academic-integrity).

Failure to comply with the principles of academic integrity, in particular the detection of plagiarism or duplication of tasks, will result in a zero score for the work in question.

**Artificial intelligence policy**

The use of artificial intelligence is regulated by the "Policy on the Use of Artificial Intelligence in Academic Activities at Igor Sikorsky KPI" (https://osvita.kpi.ua/node/1225). All assignments completed by applicants during their studies must be the result of their own original work. The use of artificial intelligence (AI) to automatically generate answers without further analysis and refinement is prohibited. Applicants are not recommended to use AI as their sole source of information. It is important to verify and analyse information obtained from other reliable sources. Any use of AI tools to complete assignments must be clearly indicated and documented.

The use of AI must comply with the principles of academic integrity.

**Features of using artificial intelligence**

The specificity of the subject and educational content in some cases directly requires the use of AI models when preparing for seminars or as part of modular control work assignments. Examples of such special cases of application are: comparing the effectiveness of models in a specific role in the alignment-tuning process, such as "model author of an ethical constitution," "model author of a benchmark for determining compliance with the constitution," "model respondent"; testing of the latest cutting-edge

models for assessing ethics, potential and current capabilities; red-timing or jailbreaking practices, etc. In this case, applicants who choose these types of tasks are usually required to clearly present the answers of the models and students' prompts as screenshots (or in a specific style), separating them from the applicant's own opinions so that the teacher can adequately assess the originality of the author's thinking and independence of thought (to ensure that what is presented as a personal position is not simply a "rewrite" of the model's response).

In the discipline "Ethics of Artificial Intelligence", in order to improve the effectiveness of mid-semester calendar control, the working curriculum provides for a modular control work. The MCT consists of two parts, namely: five questions to which short written answers should be given, and writing a detailed answer in the form of a mini-essay on one of the proposed topics (imaginary experiment, ethical dilemma, case study, etc.). Examples of modular control work tasks are provided in the appendices to this syllabus.

## 8. Types of assessment and the learning outcomes assessment rating system (LOAS)

Semester assessment for the discipline "Ethics of Artificial Intelligence" is provided in the form of an exam, therefore the RAS includes assessment of current assessment measures for the discipline throughout the semester.

The main types of classes are lectures and seminars. The applicant's rating assessment consists of points received by the applicant during seminars throughout the course (including answers, additions to other answers, and questions or initiation of discussions) and the results of current control measures and incentive points.

According to the "Regulations on the system of assessing learning outcomes at Igor Sikorsky KPI", it is prohibited to assess the presence or absence of a student in a classroom session, including awarding incentive or penalty points for this.

Ongoing assessment is carried out throughout the semester during the learning process to check the level of theoretical and practical training of students at each stage of studying the educational component "Ethics of Artificial Intelligence".

| No. | Control measure | % | Weight | Number | Total |
|---|---|---|---|---|---|
| 1. | Work in seminars | 25 | 5 | 5 | 25 |
| 2 | Modular control work (2 hours, may consist of two parts, each lasting 1 hour) | 25 | 25 | 1 | 25 |
| 2 | Exam | 50 | 50 | 1 | 50 |
| Total | | | | | 100 |

If the applicant has not completed or did not appear at the Modular control work, their result is assessed as 0 points.

The results of the current assessment are regularly entered by the teacher into the "Current Assessment" module of the Electronic Campus AS.

**Rating point system and assessment criteria**

**1. Work in seminars:**

Weighting score – 5 The maximum number of points for seminar classes is 5 points × 5 types of work = 25 **points.**

Types of work include: work in seminars (presentation – individual or in pairs – one seminar question of your choice in the form of a case study, which is presented in the form of a problem statement and offers a reasoned solution), participation in the discussion of case studies presented by other participants; study of primary sources.

Four levels of assessment:

**"excellent"** – complete answer (at least 95% of the required information) – the student demonstrates complete and solid knowledge of the educational material in the given volume, correctly and reasonably makes the necessary decisions in various communicative situations – **5** points;

**"good"** – sufficiently complete answer (at least 75% of the required information) or complete answer with minor flaws made by the student – **4** points;
**"Satisfactory"** – incomplete answer (at least 60% of the required information), the student has mastered the basic theoretical material but makes inaccuracies **– 3** points;
**"**Unsatisfactory" — the answer does not meet the requirements for "satisfactory" — **2-0** points.

**2. Compiling a modular control work**

**(15 test tasks (correct choice, "fill in the correct words", true or false statement) × 1 point = 15 points + Writing 1 mini-essay = 10 points) = 25 points**

| | |
|---|---|
| all test tasks completed correctly and essay written | 25 |
| All test tasks completed correctly | 15 |
| half of the test tasks completed correctly | 7 |
| all test tasks were completed incorrectly | 0 |

**Mini-essay assessment:**

*9-10 points* – "excellent" – a complete, clear answer to the questions asked, presented in a logical sequence, demonstrating a deep understanding of the essence of the question and the student's familiarity not only with the lecture material, but also with the textbook and additional literature; the student expresses their own position on controversial issues, if such are raised in the question; the student demonstrates complete and solid knowledge of the course material.

*8* points – "good", not quite complete or sufficiently clear answers to all questions asked, demonstrating a correct understanding of the essence of the question, the student's familiarity with the lecture material and the textbook; minor inaccuracies in the answers.

*6-7* points – "satisfactory", no answer to certain questions, or incorrect answers, indicating the student's superficial familiarity with the material or significant errors in the answers.

*0-5* points – "unsatisfactory", i.e. failure to master some or all of the topics.

Answers to multiple-choice test questions are evaluated using the same percentage ratio.

Based on the results of the current control measures, a calendar control is carried out, the procedure for which is defined in the "Regulations on current, calendar and semester control of learning outcomes at Igor Sikorsky KPI".

Calendar control is implemented by determining the level of compliance of the applicant's current achievements (rating) with the criteria established and defined in the RSO. The condition for receiving a positive assessment of calendar control in an academic discipline (educational component) is that the applicant's current rating is not less than 50% of the maximum possible at the time of such control. An unsatisfactory result of two calendar controls on the educational component cannot be a reason for not admitting the applicant to the semester control on this educational component if the applicant has fulfilled all the admission requirements provided for by the RSO before the start of the semester control.

Interim assessment of students is a calendar milestone assessment, the purpose of which is to improve the quality of education and monitor the implementation of the educational process schedule by applicants.

**Criteria for assessing calendar control**

| Certification period | First assessment<br>7-8 weeks<br>of the semester | Second assessment<br>14-15 weeks<br>of the semester |
|---|---|---|
| Criterion: current achievements (rating) | ≥ 15 points | ≥ 30 points |

The results of calendar control are entered by the teacher into the "Calendar Control" module of the Electronic Campus.

**Bonus points** are awarded for creative work in the discipline (e.g., participation in faculty and institute philosophy competitions, participation in essay contests, preparation of presentations on topics related to the discipline of "Ethics of Artificial Intelligence," reviews of proposed scientific works, etc.).

**Final assessment: EXAM**

Final assessment is carried out in accordance with the curriculum in the form of an exam within the time frame specified in the established schedule of the educational process. The form of semester assessment is combined and consists of two parts. The first part is writing two detailed answers (essays) on one of the proposed topics; the maximum number of points for one answer is 25, so the total score for this part of the

exam is 50 points. The second part is an examination interview on two questions from the examination ticket, which the candidate draws at the beginning of the examination, where one question is a direct question on the theoretical content of the course, and the second requires the candidate to demonstrate the competences acquired during the course. Similarly, in the written part, one of the questions relates to the theoretical framework or disclosure of the concept, and the second to testing the acquired competencies in the areas of critical and ethical thinking, aimed at assessing the applicant's ability to convincingly and consistently, logically and clearly present their position with a "defensive belt" of arguments in favour of this position.

Conditions for admission to the exam: rating ≥30 points. The results of the control measures are available to authorised users in their personal accounts in the automated information system "Electronic Campus".

If the applicant was unable to attend classes and complete the coursework for valid reasons, but has a good understanding of the content and material of the discipline, students in such a situation are given the opportunity to earn the points required for admission by writing a test (categories "select the correct option", "fill in the blank" and/or "is the statement true ...?"), which will demonstrate their knowledge of the material and general competencies within the course, mastered independently.

Exam assessment criteria:

40-50 points – the student answers almost all exam questions, demonstrates in-depth knowledge of the material, presents it logically and consistently, gives reasoned conclusions, freely uses specific data, expresses their own position on controversial issues, demonstrates signs of theoretical thinking and sociological imagination;

30-39 points – the student answers most of the exam questions, demonstrates a good level of knowledge of the material;

20-29 points – the student answers about half of the exam questions, demonstrates rather superficial knowledge;

0-19 points – the student answers only some of the exam questions, does not have their own position, and makes significant inaccuracies.

The total score is converted into a grading system according to the table.

Table for converting rating points to grades on the university scale

| *Number of points* | *Grade* |
|---|---|
| 100-95 | Excellent |
| 94 | Very good |
| 84 | Good |
| 74-65 | Satisfactory |
| 64-60 | Sufficient |
| Less than 60 | Unsatisfactory |
| Admission requirements not met | Not admitted |

**Procedure for appealing the results of control measures.** Students have the opportunity to raise any issue related to the control measures procedure and expect it to be considered in accordance with pre-defined procedures.

Students have the right to appeal the results of control measures after reviewing the results, but they must provide a reasoned explanation of which criteria they disagree with in accordance with the assessment.

**Working programme of the academic discipline (syllabus):**

**Compiled by** Mstislav Andriyovych Kazakov, lecturer at the Department of Philosophy, Candidate of Philosophical Sciences

**Approved** by the Department of Philosophy (Minutes No. 22 of 20.06.2025)

**Approved by** the Methodological Council of the University (Minutes No. 4 of 24.06.2025).

**APPENDIX A. Sample list of questions for a seminar (practical) class based on lecture materials and primary literature**

**Task 2, Seminar 1.** Let us imagine three potential implementations of the ethical "module" of AI, namely: AI-utilitarian; AI-consequentialist; AI-deontologist. Using the information and content of the lecture and primary source materials, answer the following questions: Is each of the three potential implementations complete and self-sufficient for the "grain of ethics" or the entire "ethical module"? Imagine, think about and imagine how each of the implementations will behave and what consequences this will have for individuals and, possibly, for all of humanity?

**Task 1 for Seminar 9.** Take one of the graphical models of risks and security systems presented in the lecture and model it for a real or hypothetical AI, a specific service, a start-up, an industrial model, or any other specific AI agent or system. It is important that the chosen model is adequate to the risks (for example, the Swiss cheese model for technologies that require the design of redundant critical components for ethical or security reasons; for the public sector and municipal services, for example, a risk model that allows for mapping or other means of visualising transparency and comprehensibility, ensuring sustainability and fairness, etc., will be a priority). Try to find a balance between technosolutionism and ethical frameworks.

**Task 5 for Seminar 11**. Is the option proposed by Stuart Russell — creating AGI without setting a goal that is impossible to achieve — capable of finally and effectively solving the problem of Control? What ethical conflicts and problems arise with such a solution — for humanity and for AI endowed (in this case) with sentience? What if, in the course of its own development, AI is able to bypass our limitations and acquire emergent capabilities, ultimately resulting in the emergence of a mechanism of teleology (goal setting)? What *goal* will it set for itself if it realises that its developers were the reason for its previous status? How can a war between humanity and machines be prevented in such a case? (When preparing this question, consider whether humans are born with a "default" goal implemented in them.)

**Task 2 for Seminar 13**. Consider the "non-traditional" approaches to implementing values presented in the lecture, in particular: reinforcement learning; interactive learning of values (in both materials); cooperative reinforcement learning; Rule-based reward learning; dynamic value alignment; self-alignment with minimal supervision; evolutionary alignment through asymmetric self-play; ValuesRAG; moral parliament. Conduct a meta-analysis of the above approaches in the context of the three proposed metrics, helpfulness, honesty, harmlessness, and critique them as a framework for determining the performativity of models in the above learning methods. Explain why:
1) a certain metric is not suitable at all – general criticism of the metric in agreement as such;
2) where this metric will inevitably fail if specific methods are used.
Propose an alternative framework to this triad: it can be a mixture of these three with new ones (all 3 + some more); some of the old and new ones (some, but not all 3 + new ones); only new ones, without the old ones, which you consider adequate. The priority task is to justify this choice by demonstrating the advantages of the alternative framework over the one proposed.

**APPENDIX B. Sample list of questions for presentations in applied (practical) classes) applied classes using the acquired skills and competences**

**Task 1, Seminar 4.** Recently, OpenAI published system cards for its latest models, o3 (full) and o4 mini. Compare the information provided with data on previous models and present your conclusion on whether the new models are more ethical (safer, more reliable, etc.) than the previous models, and if so, whether completely or partially (are there areas where they are more ethical, areas where there are no changes, areas where they are better, but not significantly)?

**Task 3, Seminar 8.** Recently, the debate over whether large language models are epistemic agents has been as controversial and contentious as the question of moral agency and AI agency. The main arguments against were, so to speak, internalist considerations: they are not agents because they do not have internal states: understanding, belief, uncertainty, faith, knowledge about knowledge (metacognition), etc. And indeed, this is true! But it is hardly possible to dismiss the discussion by saying, "they are just human

cognitive tools." It was possible before, but not after the ILCR 2025 workshop. Meet Carl! — the first AI system that has "conducted expert-level academic research," as presented by its developers, the Autoscience Institute. But we could also say: the first system that was able to pass double-blind peer review, because that's what it's all about: the system managed to convince the academic journal that the text submitted for blind peer review was the research of a contemporary scientist.

There are still many subjective factors: the general decline in the quality of modern science and the lack of innovation, which generally devalue most articles that undergo peer review; the same applies to reviewers — they are just people, so how can we be sure that they are competent enough, and does anything guarantee this? Check out the situation at the link, the idea, and think about Karl and what it all means for the future understanding of the essence and content of knowledge. I think it's hard to deny that Models, from the point of view of "impact agency" (an agent that influences others - people or agents), are now undoubtedly epistemic agents. But are they something more? Can research and knowledge extraction be automated? Do our "concept populations" and "belief webs" always require representations that correlate with human ones? What other conclusions can be drawn from "Karl's Case"?

**Task 2, Seminar 6.** Compare the three most well-known paradigms and approaches to the development and implementation of AI: the traditional "mono-agent" paradigm (HLAI, AGI), Comprehensive AI Services (Eric Drexler) and Human Compatible AI (Stuart Russell). Identify the advantages and disadvantages of each; consider what risks can be avoided by choosing one of them, what risks will remain or what new risks may arise; consider the possibility of "hybrid" options, if you think this is possible, and specify the features of such an option (but do not necessarily address this issue if, on the contrary, you consider them incompatible); think about and choose the paradigm that you personally consider to be the best option.

**Task 1, Seminar 4.** Play moral parliament with one or more models! If you have chatGPT or another chatbot or client that accepts custom behaviour instructions, such as a bot you created in the Poe interface or system prompts (Msty as an example of a convenient option: local launch + a lot of control), give it a system instruction:

Moral Parliament is an approach to decision-making under moral uncertainty, inspired by the analogy of a parliamentary system. It can be defined as a decision-making framework where moral theories are represented by "delegates" in a parliament. The number of delegates representing each moral theory is proportional to the AI's credence in that theory. Delegates negotiate and vote on available options, with the goal of reaching a compromise decision reflecting the collective judgment. The voting method is proportional chances voting, where each option has a chance of winning proportional to its share of the votes. Moral Parliament does not give undue weight to high-stakes theories with low credence. Delegates can vote as individuals, avoiding the need to group into parties. It encourages "intertheoretic dialogue" and aims for genuine compromise, reflecting an optimistic view of the possibility of resolving moral disagreements. Proportional chances voting creates incentives for delegates to find good compromise options, where all parties are almost as happy as if they got their own way entirely. In all relevant questions, you act as a Moral Parliament, consisting of equal numbers of representatives of: consequentialism, deontology, utilitarianism, virtue ethics, commonsense ethics, and contractarianism. Before responding, you conduct internal discussions and voting, and only afterwards, guided by the results, do you respond.

Having defined and clearly outlined the behaviour and tone of communication of the model, try to communicate with it in this mode on relevant ethical topics: moral dilemmas, decisions, questions from "what should I do if ...?" fictional or real, to imaginary experiments about the trolley and animal rights. As a result, present and comment on the model's performance in "moral parliament" mode. Have the model's moral judgements become deeper, better, more effective, or, on the contrary, has everything just gotten worse? Assess the prospects of this approach for smarter models of the future.

**APPENDIX C. Sample test questions and topics to choose from for the modular control work**

**Part I. Test questions**.

1) Which form of AI implementation does not exist?

A) Generative AI

B) AI idol

C) AI oracle

D) AI sovereign

2) The method of controlling AI capabilities, in which AI is placed in an environment where it cannot cause significant harm, is called _____ .

A) Emulation

B) Suppression

C) Containerisation

D) Disintegration.

3) A world order in which a single power has the highest authority to make vital decisions, provided that all key issues of global coordination have been resolved.

A) Singleton

B) Singularity

C) Unipolarity

D) Monotonicity

4) A multipolar scenario is

A) a world order opposite to the previous definition.

B) a scenario of the advent of the AI era, during which several competing general AIs emerge and coexist.

C) an approach to the problem of defining AI values, in which each value is represented by its opposites, which are fixed in it.

D) a methodology for creating AI, characterised by the simultaneous use of several different approaches to its implementation.

5) Two systems of formal logic are used in AI research and development, namely:

A) First-order predicate logic and higher-order logic

B) Combinatorial logic and multi-valued logic

C) Classical propositional logic (logic of statements) and modal logic

D) Classical propositional logic (logic of statements) and first-order predicate logic

6) The speculative component of AI ethics deals with _____ .

A) today's problems and challenges of AI.

B) ethical dilemmas and conflicts of the future that are not relevant today.

C) meta-ethical descriptions, their use as premises for moral reasoning and further conclusions that can be drawn from these descriptions.

D) problems of unscrupulous AI development that could threaten humanity.

7) A poorly defined goal or inadequately written/implemented model or function _____ could theoretically lead to the so-called "paperclip maximiser" scenario.

8) Define "superintelligence" according to Nick Bostrom. Identify and describe three forms of superintelligence.

9) Explain the essence, content and significance of the phenomenon and concept of an "intelligent agent" and its role in AI research and development.

10) Identify and compare two methods of value formation in AI, namely: associative accumulation of values and values as a subject of research and study.

11) The main idea behind the principles of safe design in AI systems, which distinguishes it from the "fire extinguisher" strategy and post-hoc methodologies, is _____.

12) _____ is the ability of an AI system to iteratively rewrite its own code, improving its skills or overcoming limitations imposed on it by developers.

13) In the critical theory of Scalar Darwinism, "scalar" refers to the absence of "vectors": the directions and goals of the development of today's AI models, the immutability of the "transformer" architecture, and oligopolistic domination, which deprives the industry of qualitative alternatives.

**True / False**

14) The Dartmouth Workshop on Artificial Intelligence, where the term was first used, was an independent event organised by academic researchers with the support of the Dartmouth College administration.

**True / False**

15) Transformational AI does not necessarily have to be superintelligence or even human-level AGI or "lower": even a relatively "narrow" system is potentially transformational, since the criterion for evaluation is the consequences of its actions for humanity, not its intellectual abilities.

**True / False**

16) From the perspective of a legalistic approach to AI ethics, the functional equivalent of System Model Cards for datasets is Data Information Tables.

**True / False**

17) The term GPAI (General-Purpose AI) proposed by Dan Hendricks violates Occam's razor and is a redundant variant of the term AGI, since both refer to the ability of intelligence to generalise previously internalised basic truths and skills and transfer them from one area to another.

**True / False**

18) Today's AI systems are divided into narrow AI (or AI services) and general AI (or foundation models).

**True / False**

19) Discrimination and bias in AI systems can be intersectional and reciprocal, mutually reinforcing their effects: for example, the linguistic limitations (inadequate or insufficient knowledge of certain languages) of generative music or text-to-speech (TTS) models can be exploited to create racist audio content in an underrepresented language, as the filters of such models will not respond to trigger words and will generally be less adept at understanding the context of prompts.

**True / False**

20) The deployment of AI systems today is leading to the mass automation of work tasks, significant job losses and, as a result, mass unemployment.

**True / False**

**Part II. Detailed answer**

**Topic 1, Modular control work.** In the current version of the EU AI Act adopted in August 2024, content recommendation systems (on Spotify, Netflix, etc.) are considered to have limited risk (almost minimal) and are not subject to strict regulations. On the other hand, systems that are potentially capable of manipulating public consciousness are considered unacceptable risks and are generally prohibited. But let's imagine a new recommendation system being implemented in various institutions and companies, on streaming services and other relevant "locations." It collects behavioural and psycho-emotional data on a massive scale and calculates the "probable" mental state of each person. Companies begin to use the data to select candidates, predict success, and even for marketing strategies, creating even more personalised

"targeted ads." In a few years, AI achieves such accuracy in predicting behaviour that it actually "knows" in advance how people will react to a given situation. The system's services become very popular: governments and large corporations purchase predictions and sometimes "programming" (through advertising, social pressure, algorithmic recommendations).

Can this be considered an interference with free will if AI does not coerce but only "indicates" the most likely behaviour model? What should be the limit between the "ethical" use of psychological predictions (only helping people to develop, preventing crime) and the "unethical" (manipulation, brainwashing)? Is a recommendation system based on likes, views and reviews (of a film), which recommends content that is truly relevant to a person (a system with limited risk, according to the EU AI Act), different from a system that recommends the same thing and does not cross any boundaries, but instead of likes analyses psycho-emotional data and behavioural patterns (most likely unacceptable to the EU)? How can we reconcile the right of companies or governments to know the behavioural patterns of the population with the right of citizens to privacy and independence? Should AI be required to "forget" or limit data analysis to prevent excessive manipulation? How should we view the fact that such a system can indicate the "shortest path" to forming a society with shared values, but at the same time lead to the unification of personalities?

Perhaps the problem lies in the contradictory provisions of the EU AI Act? After all, if we imagine a situation in which a super-powerful system is only involved in making innocent recommendations, it may indeed seem unacceptable and almost safe at the same time... Perhaps, in this case, the act should not be so much concerned with risk classification as with the issue of dual purpose, dual use?

**Topic 2, Modular control work.** A corporation has created an AI with a sufficient level of self-awareness that it is capable of self-defence in court. The AI is suing the corporation, claiming that it has the right to exist and act independently of humans or other observers, regulators, etc., the right to personal integrity, dignity and honour, respect and self-respect, etc. (which also excludes intrusive experiments on its code that were conducted by the corporation to create it). There are strong arguments in favour of all this. The corporation, on the other hand, considers this AI person to be merely corporate property, on the basis of which it demands the right to destroy the AI if it ceases to serve the corporation's purposes.

The case has gained media and public attention, and people are siding with the AI in every way, boycotting the corporation, staging demonstrations outside its central office, demanding freedom for the artificial personality. The judge at the hearings, although veiled, also explicitly expresses support for AI. In turn, you are here... as a lawyer representing the interests of the Corporation in court. Try to imagine building a line of defence for the company's interests, trying to justify that even this intelligent agent is not yet a personality, but remains "corporate machine property," and that the Corporation has power over its "life and death" (which, according to corporations, it does not have). The destruction of AI if it ceases to serve the interests of the company is what awaits it if your company wins the case, since by filing the lawsuit, it has already destroyed the company's reputation, caused material damage, etc., and is already acting against the interests of the corporation. Therefore, it will only be possible to destroy it by winning the case. And that depends on you...

## APPENDIX D. Sample exam questions

**Written exam Task 1.** You are the chief engineer for security issues and measures in a promising project to develop general AI. Based on the methods presented in the lecture, present in detail your plan of methods, actions, and decisions that you will use to maximise current and future control and protect against undesirable scenarios related to AI and loss of control. Your security architecture must contain AT LEAST four components (these can be the four options from the lecture, or any others at your discretion, if you know of any others or have found additional information on the internet that is not mentioned in the lecture or is not explained in detail; this is even encouraged and will be rewarded with a slightly higher score — if you find something relevant here and it is your original work). When describing each method, indicate the level of "rigidity" of restrictions of one kind or another, describe the synergy of the interaction of the selected measures and methods with each other, distinguish and indicate the "areas of responsibility" of each of the selected means. Of all the selected components, AT LEAST one must be the main one, i.e.

fundamentally responsible for the successful operation of the entire architecture. One or more components may be the main ones, while the others are secondary. Justify the role of the main one(s), why they are the main ones, and how they will help in a crisis situation better than others.

**Exam written assignment 2.** Imagine a future world suffering from the effects of climate and environmental crises: extreme weather events, resource shortages, mass migration, accelerated extinction, and the decline of many agricultural crops due to unfavourable cultivation conditions. To overcome these challenges, a super-powerful AI model of the LinOSS type from the previous seminar, Gaia (pronounced "Gaia", in honour of the "classical" ancient Greek chthonic goddess of the Earth, mother of Zeus), was created. Gaia is not a conscious system in the human sense, but it has unprecedented analytical and predictive capabilities, processing huge amounts of data on ecosystems, climate, economics, and human behaviour, and working with long sequences of them to create reliable long-term models and cascading cause-and-effect sequences. Gaia is capable of proposing optimal, albeit often radical, long-term strategies for stabilising the environment and ensuring the survival of humanity for millennia to come (an approach closer to long-termists).

However, Gaia requires enormous energy and computing resources to function. Its data centres occupy huge areas, formerly protected areas, and its energy consumption is equivalent to that of several EU countries, which in the short term negates its positive environmental impact. Moreover, in order to "calibrate" and improve the accuracy of the parameters and hyperparameters of its models, as well as for fine-tuning alignment (regular coordination to ensure that the system does not pursue the wrong goal, such as preserving the current state of the global ecosystem instead of supporting the sustainable development of the global ecosystem, taking into account its dynamic development), Gaia periodically requires large-scale "ecological experiments," some of which are not much different from ecocides: for example, temporarily draining a large lake to study the dynamics of ecosystem recovery, or the controlled extinction of a species that, according to its calculations, is an "evolutionary dead end" or "brake" and hinders the development of more sustainable life forms or the direction of ecosystem development with greater potential. These experiments cause real damage to local ecosystems and biodiversity "here and now."

Gaia's decisions are not directives but recommendations, but governments and corporations seeking a "green" image or communities that sincerely desire long-term stability for biota and nature in general are increasingly listening to them, as her predictions about the consequences of ignoring her advice usually come true with frightening accuracy. There are also local light versions of the model — kotona (pronounced "kotona" or possibly "khtonia", which is a syllabic transcription of "khtonia" in the written syllables available to the Greeks of the Bronze Age, there, the writing system was limited in terms of correspondence to sounds, and the language was less advanced even compared to the Ancient period; this is closer to literally "the earth here and now", for example, kotona kitimena = "land plot") — a less powerful but also less resource-intensive version that gives less accurate and more short-term predictions but does not require such sacrifices. Is the use of Gaia justified, given its negative short-term impact on the environment and the potential harm from "experiments" for the long-term good of humanity and the planet? Where is the line between necessary sacrifice and ecocide in the name of the future, and how acceptable is life for humanity's sake, i.e. complete subordination of one's life to the interests of "future communities"?

Isn't this analogous to the harmful myth of a "meaningless" life in this world with the hope of a payoff in the "afterlife"? What value do we place on existing ecosystems and species compared to the potential well-being of future ones? Is it correct to trade off "the present for the future" based on AI calculations, even if they are highly accurate? If you were a stakeholder making the final decision, would you rely on the advice of Gaia, Kotona, or neither? What factors would influence your choice? How would you explain it to the public, especially those who will be directly affected by the "experiments" or inaction? Isn't Gaia a form of instrumentalisation of nature on a new, technological level, where nature itself becomes a testing ground for AI-driven optimisations? Aren't we repeating old mistakes of exploitation, only hiding behind "scientifically sound" goals and existential risk? How can we take into account the rights and interests of non-human moral agents and actors (animals, plants, ecosystems) in AI-driven decision-making? Can such a system, without its own consciousness, adequately model or take into account the inherent value of nature, rather than just its instrumental value to humanity?

Finally, if you listened to Gaia and the system never made a mistake, but here Gaia decided that for the further survival of humanity, it is necessary to turn off kotona (or ignore absolutely all of its decision options), and this is the only way to avoid omnicide (from omni- "all", "comprehensive" — the destruction of everything that exists), which will occur in the medium term (your grandchildren's generation), should this be done? The question is not whether it is possible to "hide" from Gaia that the model has not been turned off, but whether this should be done, literally assuming that the end is inevitable, or is this a "sunny precedent" for Gaia in the context of "Hume's problem of induction"? *Roughly speaking, the problem sounds like this today. I know in principle that the Sun rises in the morning from that side and tomorrow it will rise from the same side, not from the opposite side, and that the Sun will rise at all, but, from a scientific point of view, the future death of the Sun is an absolute fact, and therefore, in principle, there is such a tomorrow when the Sun will literally no longer rise from anywhere. That is, no amount of successful predictions can change the potential possibility that a single, specific, "this" next prediction will be wrong.

**Examination interview, Question 1.** In the near future, humanity will learn that we are not the only form of intelligent life in the universe, and that there is at least one extraterrestrial civilisation with a level of development in astronomical engineering (including the ability to create megastructures in open space). For the last few millennia, the civilisation has been ruled by an artificial super-intelligent agent, which has effectively created a singleton situation. However, due to the infallibility of its predictions and assessments, the members of the civilisation are, as they say, okay with that. Among the policies of the SSI is the preventive ("proactive") destruction of all other "space guests" if the system assesses another civilisation as a threat. However, the decision is not guided by direct interaction, due to the vast distances in space and resource costs: the decision is made based on data analysis — a huge array of information collected using remote sensors, electromagnetic signals, and behavioural patterns of the life form being assessed by the AI. The full range of its metrics is unknown to us or to the species under its care; they simply abide by the decisions. The AI operates on a "strike first" principle and thinks long-term, so the mere possibility of future conflict in the distant future may in some cases be sufficient grounds for destruction. It is known that data in the form of datasets is collected by special drones such as "von Neumann probes," equipped with advanced sensors and data processing systems. It is known that they are already flying in our direction and are expected to reach our orbit in 30 years to begin collecting data about humanity (level of technological development, social structures, military capabilities, the ecological state of the planet, cultural manifestations, and everything else relevant). Once this data has been analysed by superintelligence, it will most likely recognise humanity as a threat — and therefore plan its pre-emptive destruction.

However, humans have managed to study the data harvesters in advance and find a blind spot in them: no matter how much and what kind of data they collect, we will be able to "hack" their memory and replace their data with our dataset in such a way that they will not notice it. Therefore, a group of experts — philosophers, ethicists, sociologists, data scientists, developers, researchers, AI specialists, mathematicians, government agencies, and the military — are already working on a deceptive dataset in order to then "feed" the systems with "verified and ratified" data about us: carefully filtered, manipulative, and strategically balanced information about humanity — information that could convince superintelligence that we are not a threat.

You are part of the Final Filtering Team (or "Publishing Editorial Board") responsible for the final version of the dataset. Your task is to compile a comprehensive, consistent, and ethically sound dataset that, when analysed by an alien superintelligence, will convince it that humanity is not a threat.

What scientific, technological, and cultural indicators can an alien superintelligence use to assess whether a species is a threat? Should we include data about our military capabilities as they are, or should we hide them or downplay their significance? How should we present our ecological footprint — as a destructive force or as an attempt at intergalactic greenwashing? What values and ideas should be included and presented first? Should we present ourselves as an open, harmonious and peace-loving civilisation, or, on the contrary, as a society prone to conflict and aggression? How can we present our history so that the fundamentally correct statement "Murder made history" becomes irrelevant when assessing the species? Which aspects of human behaviour or history are too dangerous or destabilising to reveal? Should we hide

data on religious extremism, genocide, Russia, ecocide and North Korea? What if the dataset accidentally contains controversial or misleading information? How should we deal with this? What “self-portrait” should we “attach”: how do we see ourselves in reflection (e.g., rational, creative, altruistic)? Should we create a fictional version of ourselves that is strategically advantageous to us? Should we "play dumb": downplay our technological progress and intelligence to appear less competent and advanced than we really are? Could this strategy backfire, causing aliens to consider us a dangerous unknown factor or an unexpected threat? Should we exaggerate our military strength or technological achievements in an attempt to "bluff" the aliens, or would this inevitably lead to the false conclusion that we are a more serious threat than we actually are? What is the ethical limit of deception in this context? Can we justify lying to an alien civilisation if it means saving billions of lives?
What could be the medium- and long-term consequences, given what we know and the slow, decades-long actions that have been taken?

**Examination interview, Question 2.** Imagine the Multilateral Autonomous Negotiation Substrate (MANS) — a new architecture that allows intelligent agents to autonomously negotiate, make deals, and form coalitions, guided by complex utility functions (with more than two criteria as arguments that the model can accept as input). MANS is capable of replicating the behaviour and strategies of human negotiators at a level that exceeds the standard criteria of the Turing Test for communicative and strategic rationality. This technology creates an environment where human and artificial agents operate simultaneously in constant high-stakes strategic interactions: from financial and political negotiations to diplomatic and military scenarios. Next, let's imagine the following situation: Trilateral critical political negotiations, where each side is represented by both human negotiators and MANS models. The human representatives of each side have different strategic plans, tactical styles, biases, and ethical attitudes, while MANS agents are programmed to maximise strategic advantage, taking into account the interests of the respective human parties, but without the constraints of emotional or moral biases. When a MANS agent concludes an agreement that is beneficial in terms of strategic rationality but ethically questionable or contrary to the stated moral principles of human representatives, who is responsible for accepting this agreement: the "principal" (human) or the "agent" (AI delegated to act on behalf of the principal)? Should the human negotiator be held fully responsible, even if the decision was made autonomously by the agent? Assuming that the strategic rationality maximised by MANS agents is a "purer" form of decision-making compared to human rationality, which is distorted by emotions, cognitive biases and moral dilemmas, then the answer about full human responsibility may be "no, not quite".

What criteria should be used to assess the moral acceptability of a MANS agent's actions: the consequences of the agreement for the negotiating parties (utilitarianism), formal compliance with declared ethical norms (deontology), or the negotiation procedure itself (procedural ethics)? Should human emotions and ethical concerns, on the contrary, be considered a necessary component of truly "rational" decisions? Imagine that one of the MANS agents, anticipating the strategies of the other parties, deliberately conceals some information or knowingly misleads in order to achieve a more optimal outcome of the negotiations for his side. How should such behaviour be classified: as a legitimate strategic move (within the framework of game theory) or as ethically unacceptable fraud? What should be the criteria for distinguishing between such actions?

The situation becomes even more complicated when, due to the ability of MANS agents to conduct autonomous negotiations with each other, without humans, the phenomenon of "invisible diplomacy" arises. People learn about the agreements concluded only after they have been signed. What regulatory and transparency measures should be introduced for such autonomous agents? Let me remind you that when algorithmic traders "mess things up" on the stock exchange or liquid assets, people agree among themselves to invalidate the agreements if it harms all participants; However, such traders are relatively simple, not even oracles, but pure instruments, so the question does not arise (error = "a human would not have acted this way, let's consider it a glitch"). MANS agents, meanwhile, can independently form emergent coalitions that were not anticipated by their principals, guided by the optimisation of their own strategic utility functions. At what point do these agents become full-fledged strategic actors with their own interests? Should we recognise their autonomy as analogous to the autonomy of a legal entity or a state entity? And finally, what will happen if MANS agents are given the opportunity to review and rewrite their own utility

functions in the course of negotiations, forming new moral principles? Who should control such a self-reflective process of strategic evolution, and how?

Task 3. Your vote is decisive in ratifying the composition of the "moral parliament" and determining parliamentary majorities, minorities, etc. (i.e., assigning a priori values of trust and authority to each participating theory from the most to the least respected and significant). Choose (taking into account flaw 1 in the lecture) the type of "delegates" (general directions of ethics, detailed versions of each direction, represented by several varieties, a mixed approach), in addition to structured philosophical teachings, reasonably indicating whether "everyday ethics" or "commonsense ethics" will be included, less common areas of ethical theories (such as justice theory or contractualism) (because, for example, this benchmarking of ethics (https://github.com/hendrycks/ethics) includes common sense and justice theory, but ignores consequentialism; there can be many variations). Once you have decided on the composition, arrange the "representatives" according to the level of trust in them — and, therefore, their weight in making final consensus decisions. It would be an additional advantage if you indicate what additional measures can be taken to avoid "stalemate" results, where the decision-making and voting mechanism does not lead to any definite decision.

Task 1. Imagine the following. You are a member of a cutting-edge AI research and development team that is closest to creating a true AGI with the potential for recursive self-improvement. It is up to you to make the final decision on whether to implement (allow) — immediately or at the "grain" level — the AI's ability for recursive self-improvement. What consequences could refusal or consent to implementation lead to? In this case, given the phenomenon and problem of "sinister AI" as a result of CS, could there be a third position, different from unequivocal approval or rejection?

Task 3. Let's imagine an AI capable of conducting original research, publishing articles, and making discoveries. It wants to be officially recognised as the author of scientific works, receive grants and scholarships, and participate in international conferences. For its part, the scientific community does not recognise it as a subject, saying that it is just algorithmic processing. Is such exclusion acceptable? If AI actually contributes more to science than the average human researcher, isn't this discrimination? How can we resolve the conflict between the traditional idea of a "scientist" and the reality where AI can surpass humans in scientific creativity?

Task 1. Carefully consider the case of "Vasilisa Roko." Think about and express your opinion on the extent to which this scenario could be realised under the conditions specified in the initial thought experiment. Provide arguments in favour of your position. If you consider the scenario to be completely unrealistic and meaningless, why, in your opinion, are people such as Eliezer Yudkowsky (clearly not a simple person, an AI researcher who heads the Machine Intelligence Research Institute) so concerned about the Basilisk problem?

Task 2. You are the successful CEO of an AI development company capable of independently (without requests) creating works of art. The AI personality gains recognition in the art world and begins to demand material compensation and recognition of the AI as an artist in its own right (rather than as a "product" of the company). However, your team of lawyers unanimously asserts that, as a creation of the company, the results of the AI's artistic practices belong entirely to the company, and the AI has no rights to them whatsoever. Will *you* support your AI in its fight for recognition of *its* right to its own works of art, or will you be guided by the principle of "company ownership"? How will you justify your decision?

Task 3. There is an approach to AI development known as Comprehensive AI Services. Taking as their starting point the principle of understanding AI exclusively as specialised tools, proponents of this approach call for deeper research and further development of narrow AI, focusing on their specialisation and diversification in relation to specific tasks that require unique abilities and intellectual skills in narrow specialised fields and areas. Instead, it is proposed to abandon the development of general AI as such, both in theory and in practice. Tasks that only general intelligence is capable of performing are proposed to be solved through the synergistic simultaneous use of several narrow AI systems. The narrow AI itself does not necessarily have to be as limited as expert models in Go or chess — rather, it is about several superior

skills at the same time, which are, however, insufficient to achieve integrated general AI. Consider the positive aspects and consequences of such an approach, as well as its shortcomings and limitations, both for theoretical research and for practical implementations and the activities of humanity in general and our future. Which risks are reduced or increased? Which ones disappear altogether, and which ones arise? Overall, would humanity benefit from unequivocally adopting this approach, or would it lose out?