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

Кафедра філософії

ЕТИКА ШТУЧНОГО ІНТЕЛЕКТУ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	Другий (освітньо-науковий), магістр
Галузь знань	С - соціальні науки, журналістика, інформація та міжнародні відносини
Спеціальність	С5 Соціологія
Освітня програма	Аналітика соціальних даних
Статус дисципліни	Вибіркова
Форма навчання	Очна (денна)
Рік підготовки, семестр	1 рік, 2 семестр
Кількість кредитів ЄКТС	150 годин (30 лекційних, 30 практичних, 90 СРС)
Семестровий контроль/ контрольні заходи	модульна контрольна робота, екзамен
Розклад занять	https://schedule.kpi.ua/
Мова викладання	Українська / Англійська
Інформація про керівника курсу / викладачів	Лектор, Практичні / Семінарські: Казаков Мстислав Андрійович
Розміщення курсу	Платформа дистанційного навчання Сікорський / Google classroom; посилання

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчання та результати навчання

За останні 15 років, найбільший прогрес у big tech демонструють дослідження та розробки в області машинного навчання, більш відомої сьогодні як «штучний інтелект» - сукупність технік, технологій, методів створення та розгортання «розумних» машин та програм – тобто, таких, що частково відтворюють ту чи іншу здатність або активну властивість інтелекту. Комп'ютерний зір, розпізнавання зображень та візуальне міркування, обробка природної мови, генеративні алгоритми – основні області машинного навчання, у яких прогрес є найбільш відчутним, виступаючи передумовою виникнення фундаційних моделей, так званих GPAIs (Систем Штучного інтелекту загального призначення). Завдяки ним, зокрема, через розгортку на інфраструктурному та публічному рівні Великих мовних моделей, ШІ сьогодні став не тільки масовим явищем, а й комерціалізувався. Все ще нескінченно далекі від так званого «Загального інтелекту», синтетичні сутності дуже добре навчилися емулювати саму антропоморфність – не лише зовнішню форму, а й «внутрішній світ» людини: думку, процес розмірковування та висновку, ментальні стани (переконання, знання, впевненості чи сумнівів), почуття, емоції; найвідомішою та найбільш затребуваною сьогодні областю емуляцій, безумовно, є продукти інтелектуальної активності, творчість в її найширшому розумінні: від наукових досліджень до музичних композицій. Впровадження систем ШІ в повсякденне життя, професійну діяльність, логістику, виробництво, адміністративні, фінансові та юридичні процеси є неминучим і обіцяє людству економічні вигоди, покращення рівня життя, вивільнення часу та свободу від рутинних завдань; в більш довгостроковій та уявній перспективі, технологія ШІ обіцяє нам вирішення глобальних проблем та створення нової,

небіологічної форми розумного життя, «сентіонавтів» відомих сьогодні як AGI – Artificial General Intelligence.

Проте є також фактом і те, що ці скалярні амбіції, мрії зазнають скалярного *колапсу*, а обіцяний золотий вік завжди відкладається на невизначене «завтра». Окрім відсутності технічних (інженерних) рішень, технологічного «лагу» (повільного та нерівномірного розповсюдження автоматизації системами ШІ), і на відміну від інших технологій із трансформаційним потенціалом (на кшталт bitcoin), ШІ є найбільш проблемною технологією з точки зору *етики*. Етичні колізії штучного інтелекту охоплюють широке коло проблем – від екзистенційних (спекулятивних) ризиків, втрати когнітивної автономії та безробіття до централізації благ та засобів, посилення нерівностей, укорінення існуючих структур влади, систем диспицлінарних практик, примусової «нормалізації», автоматизації озброювання (weaponization), контролю, нагляду за населенням, аж до використання алгоритмів як кінцевих агентів у прийнятті рішень, які напряму торкаються інтересів людини, її свободи, автономії, здоров'я та життя. Окрім людей, ШІ на сьогоднішньому етапі історичного розвитку наносить серйозну шкоду також і довкіллю, призводить до надмірних ресурсних витрат, має глибокий вуглецевий слід: навчання однієї Великої мовної моделі за кількістю тон викидів CO_2 давно лишило позаду авіапромисловість. Системи ШІ – не спекулятивна загроза майбутнього, а різномасштабні серії структурних та системних проблем, вирішення яких потрібне вже сьогодні.

Мета вивчення дисципліни – демістифікувати та демофологізувати штучний інтелект, надавши критичний інструментарій та критико-теоретичний фреймворк, методологію, техніко-технологічне та етичне знання, які для цього необхідні, сформувавши зважений та синоптичний погляд, відношення, навички необхідні для взаємодії з ним; навички такого роду є необхідними як для користувачів, так і для інших груп стейкхолдерів, позаяк джерелами етичних проблем є всі ми: академічні дослідники, розробники, власники, поставники інфраструктури, інвестори, зацікавлені групи (від маргіналізованих спільнот до «середньостатистичних» споживачок та споживачів). Свої власні інтереси переслідують тут національні держави, корпорації, військові, перетворюючи ШІ в інструмент геополітичного впливу та провокуючи гонку озброєнь, результатом якої, у «сплаві» з неоліберальним «моментом» сьогодення став поточний стан справ, який ми маємо у ШІ, такий собі «скалярний дарвінізм», гонка розробників у постійному збільшенні своїх моделей-трансформерів без цілей та інфраструктурних змін.

Починаючи від фундаментальної деконструкції значення концептів «штучний» та «інтелект» у гуманістичному, технологічному та науковому спектрах, побічно оглядаючи також деякі аспекти та проблеми комп'ютерних наук та розробок, філософії, теології, історії, літератури та прикладних промисловості та виробництва, в ході курсу ми з'ясуємо, що «етичного ШІ» ніколи не існувало, розвіявши популярний міф про «вчених-мрійників, які вирішили навчити машин думати як люди»: від кібернетики та Алана Тюринга до сьогодення, Штучний інтелект був і залишається результатом синергії на стику інтересів, законодавчих та матеріальних можливостей та амбіцій академії, військових, спецслужб, національних урядів та капіталістичних агентів, від венчурних фондів до мегакорпорацій, левіафанів так званих big tech. Але якщо етичного ШІ не існувало – ми маємо створити його! Відповідальний та людино-центрований (не антропоцентристський) ШІ є пріоритетом як для нас, так і для майбутніх поколінь людей і не-людей, перед якими ми також маємо відповідальність. Ми є основною проблемою сучасних етичних викликів, згенерованих цією технологією. Але саме ми здатні і відповісти на ці виклики, вирішити проблеми на рівні управлінських (governance), дизайнерських та розробницьких, інтелектуальних та споживацьких рішень, створити екосистему сталого та справедливого ШІ. Додатковою метою є культивація у слухачів та слухачок курсу гнучкого та адаптивного мислення при розгляді різного роду проблемних аспектів комп'ютаційної архітектури та шляхів імплементації ШІ, сприймаючи логічну основу прийняття рішень ШІ, інкорпоруючи отримані знання та навички у фреймворк соціальної, громадської та етичної відповідальності через спектр варіантів дій, вчинків, рішень, підходів, тактик та стратегій, між якими необхідно маневрувати в сучасному ландшафті ШІ аби змінити його на краще.

Після вивчення навчальної дисципліни студенти мають можливість розвинути такі програмні результати навчання:

широке та неупереджене розуміння реального стану справ у індустрії сучасного ШІ із знанням історії та генеалогії – витоків сучасних систем ШІ;

знання реальної історії ШІ – з одного боку, як предмету науково-фантастичних та філософських роздумів та історій, з іншого – як «плід» мілітарно-промислового комплексу у взаємодії з ученими.
Набуття здібностей та практичних умінь до прагматичних імплікацій ШІ у таких критично важливих сьогодні областях та сегментах людської реальності, як автоматизація, право та законодавство, воєнні дії, логістика, транспорт, промисловість, освіта, мистецтво, політика та управління ШІ загалом;
вміння генерувати нові знання, ідеї, моделі та гіпотези щодо довгострокових та високоабстрактних моральних питань та дилем;
здатність до спекулятивного та імплікативного мислення як форм випереджаючого відображення (антиципації) майбутніх етичних дилем;
включати імплікації ШІ до людського сприйняття та особистісності, на рівні буденних та нетривіальних компутаційних інтуїцій щодо оперування ШІ як спеціалістами, так і тими, хто не є фахівцями у комп'ютерних науках;
підвищення компетентності та орієнтованості в літературі, моральній філософії, менеджменті, історії;
вміння проводити, організовувати теоретичні дебати, пленарні групи та робочі групи для вирішення практичних, наявних сьогодні питань, пов'язаних із різними імплементаціями «вузького ШІ» відповідно до реальних викликів, котрі породжує його експлуатація у соціальній та індивідуальній сферах;
широка обізнаність, знання, навички та вміння до участі в спекулятивних дискусіях та дослідженнях на предмет майбутніх імплементацій ШІ та виходу за межі «вузького» ШІ до Загального ШІ та Суперінтелекту – і як екзистенційного ризику, і як потенційного вирішення всіх або частини глобальних проблем людства загалом;
критичне мислення в умовах радикальної невизначеності або відсутності критично важливих фрагментів інформації, викликаних непередбаченістю – темпоральною та квалітативною – розвитку технологій штучного інтелекту та емерджентним характером їх практичного застосування.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Пререквізити: розуміння базових філософських концептів та теорій в області етики, епістемології та метафізики; бажаним є розуміння основ комп'ютерних наук та досліджень в області ШІ (машинне навчання, структури даних, нейронні мережі, прикладний ШІ тощо) – не є обов'язковим; перевагою буде також розуміння ключових концептів соціальних та політичних теорій, таких як справедливість, публічна політика, управлінство, менеджмент тощо. Знання англійської мови є ще однією додатковою перевагою, оскільки дуже велика кількість досліджень та найостанніших новин з питань етики ШІ та розвитку індустрії загалом головним чином представлені англійською мовою.
Постреквізити: специфічні етичні проблеми (приватність, похибки, агентичність, автономія, управління ШІ технологіями, регуляція тощо); ШІ-політика та менеджмент, управлінські фреймворки міжнародного та державного рівнів; дослідницькі методи в області ШІ, аналіз даних, квантитативні та квалітативні дослідження, міждисциплінарні та трансдисциплінарні дослідження.

3. Зміст навчальної дисципліни

Тема 1. Штучний інтелект: історія та форми існування. Етика: філософська наука про системи моралі, вчинки та обов'язок. Історія та генеалогія етики ШІ.
Тема 2. Теоретичне підґрунтя етичних проблем ШІ-технологій з точки зору комп'ютерних наук.
Тема 3. Скалярний дарвінізм як логіка розвитку Штучного інтелекту в добу капіталістичного реалізму.
Тема 4. Моральний та епістемічних статус штучних агентів. Проблема пробілу у відповідальності.
Тема 5. Атлас ризиків та загроз з боку ШІ.
Тема 6. Приватність та етика в науках про дані та Big data.
Тема 7. Протидія загрозам та підходи до вирішення етичних проблем.
Тема 8. Парадигми етичного ШІ: від публічного до пояснюваного.

Тема 9. Штучний Інтелект як джерело глобальних асиметрій та нерівностей: соціальних, освітніх, економічних, епістемічних, когнітивних, кваліфікації робочої сили на ринках праці
Тема 10. Принципи безпечного дизайну систем ШІ
Тема 11. Проблема формування та імплементації цінностей (Alignment problem).
Тема 12. ШІ та Проблема контролю.
Тема 13. Штучний інтелект, екологія, лонгтермізм (парадигма «Зеленого» ШІ).

4. Навчальні матеріали та ресурси

Наведено рекомендовані навчальні матеріали та ресурси для засвоєння матеріалу, розглядуваного на лекційних заняттях та практичних заняттях. Відповідно до рекомендацій створення «гнучких освітніх архітектур», з метою досягнення дидактичних та доступних освітніх цілей шляхом включення змішаних та цифрових матеріалів, здатних полегшити процес засвоєння матеріалу слухачами не-спеціалістами в області ІСТ (інформаційно-комп'ютерні технології), до базових та допоміжних літературних джерел додається список рекомендованих відеоматеріалів (фільми, серіали, лекції, документальні фільми, youtube-контент, матеріали інтернет-архіву та відео з ліцензією Creative commons, записи вебінарів та / або конференцій), комп'ютерних ігор, чатботів ШІ та іміджбордів, релевантних до тематики курсу.

Література
Базова

1. Авдєєва Т. (2024) Дії та Мрії: штучний інтелект у публічному секторі - ГО "Лабораторія цифрової безпеки". – https://dslua.org/wp-content/uploads/2024/02/Mrii_ta_dii_kopiia.pdf

2. Бостром Н. (2020) Суперінтелект. Стратегії і небезпеки розвитку розумних машин. Київ: Наш Формат

3. Девенпорт, Т., Корбі Дж. (2018) Вакансія: людина. Як не залишитися без роботи в добу штучного інтелекту. К.: Наш формат

4. Когут Ю.І. (2024) Штучний інтелект і безпека: практичний посібник. Київ: Консалтингова компанія «СІДКОН»; ВД Дакор

5. Коцовський В.М. (2016) Методи та Системи Штучного інтелекту. Конспект лекцій. – https://shorturl.at/E16cf

6. Лі, К.-Ф. (2020) Наддержави штучного інтелекту: Китай, Кремнієва долина і новий світовий лад. Київ: Форс Україна.

7. Ліч, Н. (2024). Архітектура в добу штучного інтелекту. Вступ до ШІ для архітекторів. ArtHuss

8. Методи та системи штучного інтелекту: навч. посіб. / укл. Д.В. Лубко, С.В. Шаров. – Мелітополь: ФОП Однорог Т.В., 2019. – http://elar.tsatu.edu.ua/bitstream/123456789/15462/1/5_lubko_metody_2019.pdf

9. Рассел С. (2020) Сумісний з людиною: Штучний інтелект і проблема контролю. Київ: Форс Україна

10. Сейтуа де, М. (2023) Код творчості. Як штучний інтелект учиться писати, малювати, думати. Київ: Arthuss

11. Уолш, Т. (2025). Підробка: штучний інтелект у світі людей. Харків: ВД «Фабула»

12. Шаров С. (2023) Сучасний стан розвитку штучного інтелекту та напрямки його використання. - https://www.researchgate.net/publication/370595289_Sucasnij_stan_rozvitku_stucnogo_intelektu_ta_napramki_jogo_vikoristanna

Допоміжна

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

Література зі списку, котра може бути відсутньою у вільному доступі для завантаження на момент запиту, буде надана викладачем. Частина завдань до семінарських занять базується на перекладених з англійської та адаптованих до навчального процесу матеріалах.

Додатковий контент (відеоматеріали)

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

Додатковий контент
(інтерактивні матеріали, репозитарії, форуми, комп'ютерні ігри)

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, 08 Apr. 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/

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Навчальна дисципліна охоплює 30 годин лекцій та 30 годин практичних занять, а також виконання модульної контрольної роботи (МКР). Практичні заняття з дисципліни проводяться з метою закріплення теоретичних положень навчальної дисципліни і набуття здобувачами умінь і досвіду оперувати сучасними поняттями в галузі області етики штучного інтелекту. Виходячи з розподілу часу на вивчення дисципліни, рекомендується чотирнадцять практичних занять (з врахуванням часу на МКР).

Виходячи з принципів побудови гнучкої освітньої архітектури, методи та форми навчання включають не лише традиційні університетські лекції та семінарські зайняття, а також елементи роботи в командах та групових дискусій. Застосовуються стратегії активного навчання, які

визначаються такими методами та технологіями: методи проблемного навчання (дослідницький метод); особистісно-орієнтовані технології, візуалізація та інформаційно-комунікаційні технології, зокрема електронні презентації для лекційних занять. Комунікація з викладачем будується за допомогою використання інформаційної системи «Електронний кампус», платформи дистанційного навчання «Сікорський», а також такими інструментами комунікації, як електронна пошта і Telegram.

Етичні, суспільні та культурні імплікації штучного інтелекту не обмежуються книгами, статтями та контентом медіа-новин, а проявляються вздовж усього спектру сучасної цифрової культури: в іграх, інтерактивних симуляціях онлайн форумах та відео-ессе. З урахуванням цього, курс пропонує у якості додаткових джерел інформації теоретичного та практично-прикладного типу ряд нетрадиційних ресурсів, до яких відносяться: фільми, комп'ютерні ігри, інтерактивні інструменти, дискусії на форумах та сабреддіти, цифрові артефакти, які можуть функціонувати як повноцінні джерела для аналізу феноменологічного та опосередкованого (mediated) досвіду взаємодії або інтерпретацій ШІ.

Лекції

Лекція 1. Штучний інтелект як проблема етики

1. Етика: філософське дослідження моралі, вчинків та форм життя.
2. Штучний інтелект: загальна перспектива.
3. «Почнемо з того, що ми ніколи не були етичними» – ШІ як «міф про даність».
4. Етика ШІ, прикладна та спекулятивна.
5. Від «Ереуона» та «Дюни» до "Westworld" та "Person of Interest": спекулятивна наукова фантастика як джерело генези етики ШІ.

Лекція 2. «Скалярний дарвінізм» як логіка розвитку ШІ-технологій

1. Великі мовні моделі та акселераціонізм
2. Наглядовий капіталізм
3. ШІ-Капіталізм
4. Капіталістичний реалізм, неоліберальний «момент», технології ШІ
5. Скалярний дарвінізм як логіка розвитку та форма існування сучасного штучного інтелекту

Лекція 3. Теоретичне підґрунтя етичних викликів ШІ

1. Основні поняття та предметні області етики штучного інтелекту
2. Імплікації комп'ютерних наук: етичний вимір
2.1. Теорема про відсутність безкоштовних сніданків у пошуку та оптимізації та алгоритмічні упередження
2.2. Теорема про неможливість машинної справедливості (Machine Fairness): всі критерії не узгодити, всі побажання не врахувати

Лекція 4. Штучні інтелекти: реальні та можливі
1. Рівні інтелектуальності від вузького ШІ та мікробних колоній до суперінтелектів
2. Функціональні реалізації систем ШІ
3. Типологія інтелектуальних агентів як проблема етики
4. Етичні наслідки субстанційних реалізацій (інструменти, джини, оракули, суверени)

Лекція 5. Штучні моральні агенти та пробіл у відповідальності.

1. Чи може машина бути моральним агентом?
2. Чотири «пробіли»: проблема відповідальності
3. Проблема «Принципала – Агента» (Першого та Другого порядків).
4. Консеквенціоналізм, деонтологія, та метаетика про агентичність та агенційність (agenticness and agency) ШІ

Лекція 6. Машинна етика доби Великих мовних моделей

1. Чи є ВММ агентами: «Стохастичні папуги» або «Кандидати в сентієнти»?
2. Дефляціонізм та інфляціонізм, від редукціонізму до всеінклюзивної епістемології.
3. Емерджентні властивості, «пояснюваність» (explainability) «чорних ящиків» як поле сучасних спекулятивно-філософських дебатів
4. Емпіричний поворот у філософії ШІ: дослідження з інтерпретованості (interpretability) як доказова база нових аргументів
5. Комп'ютаційний інференціалізм: ВММ – агенти без свідомості в просторі причин

Лекція 7. Ризики та виклики штучного інтелекту.

1. Класифікуючи ризики штучного інтелекту
2. Загальна типологія.
2.1. Дискримінація та токсичний контент
2.2. Збереження приватності та безпека
2.3. Оманлива інформація
2.4. AI-Darkside
2.5. Ризики взаємодії
2.6. Суспільство, економіка, політика, екологія.
2.7. Безпека, системні збої, внутрішні обмеження систем ШІ.

Лекція 8. Приватність та справедливість

1. Етика для data-science
2. Зброя масового математичного знищення (Weapons of Math Destruction)
3. Датасети та бенчмаркінг: від вузьколобих людей до «вузьколобості за дизайном»
4. «Наддержави» ШІ та глобальна нерівність

Лекція 9. Нормативне регулювання штучного інтелекту в світі та державні стратегії AI-Governance
1.1. ЄС
1.2. США
1.3. Великобританія
1.4. Китай
1.5. Південно-Східна Азія
1.6. Африка
1.7. ООН, G7 та інші наддержавні актори
1.8. Регулювання ШІ в Україні

Лекція 10. Методи протидії етичним ризикам та викликам
1. Інституційні методи
2. Технічні методи
3. Трансдисциплінарні методи

Лекція 11. Роблячи ШІ Етичним
1. Чесність (AI Fairness)
2. Прозорість (Transparency)
3. Відстежуваність (Traceability)
4. Підзвітність
5. Надійність (Robustness)
6. Стійкість до атак
7. Ступінь довіри та узгодженість (trustworthiness and alignment)
8. Сталий розвиток

9. Свобода та автономія

Лекція 12. Публічний ШІ: Перспективи, цілі, перешкоди альтернативі пропрієтарності

1. Екосистема публічного ШІ
2. Загальні цілі публічних ШІ-проектів
3. Приклади втілення цілей: Case studies
4. Напрями розвитку публічних альтернатив
5. Ролі та рекомендації: Що *реально* можуть стейкголдери
6. Інструменти: від стандартизації до регуляції

Лекція 13. Принципи безпечного дизайну систем ШІ

1. Дизайн як метаетичний фреймворк.
2. Принципи безпечного дизайну та їх втілення:
2.1. Резервування
2.2. Прозорість
2.3. Розподіл обов'язків
2.4. Мінімізація привілеїв
2.5. Відмовостійкість
2.6. Антикрихкість
2.7. Негативний зворотний зв'язок
2.8. Глибинний захист
2.9. Приписуваність
2.10. Переривність
2.11. Агентична розбірливість
2.12. Методи контролю над здібностями
2.13. Одомашнювання
3. Три синергії: Принципи в дії.

Лекція 14. Проблема Узгодження: Цінності в системах ШІ та шляхи імплементації

1. Проблема імплементації цінностей.
2. Метрики: Корисність, Чесність, Нешкідливість
3. Методи імплементації цінностей
3.1. Нормативне програмування та його критика
3.2. Навчання з підкріпленням: від людей до моделей
3.3. Два шляхи інтерактивного вивчення цінностей
3.4. Новітні способи узгодження (еволюція, аугментація, самогра)
3.5. Моральний парламент
4. Альтернативні фреймворки: перспективи майбутнього

Лекція 15. ШІ в контексті Етики довкілля

1. Етика навколишнього середовища та Дилема лонгтермізму
2. Технологія як медіатор відносин «Людина – Природа». ШІ як Технологія.
3. ШІ як фактор екологічної деградації
3.1. Енергоспоживання
3.2. Вуглецевий слід
3.3. Водні ресурси
3.4. Екстракція ресурсів та виробництво апаратного забезпечення
3.5. Неекологічна логістика
3.6. Відбиток «Інтернету Речей»
3.7. E-Waste

3.8. Непрямі впливи
4. Наша відповідальність в епоху Антропоцену
5. На шляху до «Зеленого» (Стійкого) ШІ

Практичні (семінарські) заняття

Основним завданням циклу семінарських занять є поглиблення знань, які студенти отримують на лекціях, навичок роботи із базовою та додатковою літературою, формування вмінь аргументовано доводити власні думки, а також розвиток комунікативних здібностей. Семінарські заняття мають сприяти кращому засвоєнню теоретичного матеріалу з курсу «Етика Штучного Інтелекту». З метою інтеграції теоретичної компоненти (знання) із практичною (застосування знань) питаннями до семінару є: 1) теоретичні питання релевантні до *матеріалу* лекції, який передував семінарським заняттям; 2) самостійні дослідження case-studies за *темою* лекційного матеріалу, яку він охоплює, або за *зразком* такого роду дослідження, представленого під час відповідної лекції *або* у формі зразка-шаблона безпосередньо доданого до умов завдання; 3) кейси за тематикою лекцій, за якими здобувачі готують та презентують під час заняття доповіді, у якій представляють як своє бачення етичної проблеми зі сфери ШІ, так і варіант вирішення, який вони вважають найкращим із можливих. Для деяких вибіркових тем передбачена командна робота у групах 2-3 осіб. Розподіл відбувається наступним чином: 7 семінарів присвячено розгляду теоретичних питань, висвітлених в лекціях протягом року, 8 семінарів присвячено презентаціям кейсів та іншим формам практичної активності.

Семінар 1. Генеалогія та історія штучного інтелекту.
Семінар 2. Теореми комп'ютерних наук та машинна етика
Семінар 3. «Скалярний дарвінізм» сучасних систем ШІ та питання альтернатив
Семінар 4. ШІ, чотири Пробіли відповідальності та дві Проблеми «агента – принципала»
Семінар 5. Сучасні Великі мовні моделі як предмет етичних досліджень
Семінар 6. Ризики та загрози інтелектуальних агентів: *від* майбутнього *до* сьогодні
Семінар 7. Нормативний, управлінський та легалістський підходи до вирішення етичних проблем
Семінар 8. Інженерні рішення та імплементація бажаних якостей
Семінар 9. Публічний ШІ: Спільне як альтернатива
Семінар 10. Моделювання ризиків: Етика ШІ через призму управління ризиками
Семінар 11. Принципи безпечного дизайну: концепт, імплементації, синергія
Семінар 12. Проблеми Контролю, Узгодження та Імплементації цінностей
Семінар 13. Приватність та етичні колізії data-science
Семінар 14. ШІ та Етика довкілля в добу Антропоцену. Симуляція «Вирішальних дебатів»: Самміт стейкхолдерів, що формує всі аспекти повістки дня щодо ШІ на найближчі 20 років, від регіонального та міжнародного законодавства, до повноважень аудиторів приватних пропрієтарних технологій та напрямків досліджень, що будуть / перестануть фінансуватися.
Семінар 15. Модульна контрольна робота

6. Самостійна робота здобувача вищої освіти

Самостійна робота здобувача ВО в рамках освітнього компоненту включає комплекс тематичних питань для роздумів і практичних завдань, спрямованих на самоконтроль знань та організацію самопідготовки здобувачів в рамках кожного з практичних/семінарських занять і передбачає: підготовку до аудиторних занять із теоретичною складовою; підготовку дослідження та презентації кейсів (практичної складової); підготовку до модульної контрольної роботи; підготовку до екзамену; поглиблену підготовку до занять – опрацювання додаткової літератури та альтернативних першоджерел, наведених у додаткових списках (опціонально); підготовку до роботи в групах (опціонально).

№ з/п	Самостійна робота студентів	Кількість годин
1	Підготовка до аудиторних занять	56
2	Підготовка до складання модульної контрольної роботи	4
3	Підготовка до екзамену	30
	Всього	90

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Правила відвідування занять

Відвідування лекцій є обов'язковим. Здобувач вищої освіти, який пропускає лекції, може мати труднощі з належною підготовкою до семінарів, однак йому не потрібно відпрацьовувати пропущені лекції. Здобувачу вищої школи під час лекції рекомендується конспектувати основні аспекти, ключові поняття, визначення, класифікації та алгоритми, що пояснює викладач.

Активна участь у семінарських (практичних) заняттях є обов'язковою і має важливе значення для формування рейтингу здобувача вищої освіти. Готуючись до семінарського заняття здобувач вищої школи має обов’язково опрацювати лекційний матеріал певної теми, а також опрацювати інформацію теми, що представлена у основному списку літератури та проблемні кейси, які слідують за лекцією і виносяться на відповідні практичні заняття у якості тем на вибір для презентацій відповідей. У разі виникнення запитань або незрозумілих моментів слід обов’язково обговорити їх з викладачем. Якщо здобувач вищої школи не встиг підготуватися, йому варто уважно слухати виступи інших і намагатися компенсувати недостатню підготовку, засвоюючи нову інформацію.

Якщо здобувач пропускає семінари чи контрольні заходи з поважних причин (наприклад, через хворобу чи інші важливі обставини), він матиме змогу виконати завдання впродовж наступного тижня. Перевірка знань здобувачів вищої школи із тем, які вони пропустили, здійснюватиметься через консультації з викладачем, графік яких доступний на сайті кафедри філософії.

Бали за фактичну присутність на лекціях і семінарах не нараховуються.

Дистанційне навчання

У разі запровадження дистанційного (змішаного) формату навчання організація освітнього процесу здійснюється відповідно до Положення про дистанційне навчання в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/index.php/node/188), Регламенту проведення семестрового контролю в дистанційному режимі (https://osvita.kpi.ua/node/148).

Організація освітнього процесу здійснюється з використанням технологій дистанційного навчання, зокрема через Платформу дистанційного навчання «Сікорський» (https://www.sikorsky-distance.org) та АС «Електронний кампус» (https://ecampus.kpi.ua). Здобувачі вищої школи приєднуються до платформи «Сікорський» (Google Classroom) через корпоративну електронну пошту у домені @lll.kpi.ua.

Освітній процес у дистанційному режимі здійснюється відповідно до затвердженого розкладу навчальних занять. У режимі дистанційного навчання заняття відбуваються у вигляді онлайн-конференції на платформі Zoom. Посилання на конференцію надається на початку семестру.

Заняття в режимі дистанційного навчання проводяться через онлайн-конференції на платформі Zoom. Результати оцінювання висвітлюють у АС «Електронний кампус» на особистій сторінці здобувача вищої освіти (https://ecampus.kpi.ua).

Правила поведінки на заняттях

На заняттях слід дотримуватись норм етичної поведінки, що визначені у Кодексі честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (https://kpi.ua/code), а також Положенні про комісію з питань етики та академічної чесності НТУУ «КПІ» (https://data.kpi.ua/sites/default/files/files/2015_1-140a1.pdf).

На території університету здобувачі вищої освіти зобов'язані дотримуватись установлених Правил внутрішнього розпорядку (https://kpi.ua/admin-rule). Під час лекційних і практичних занять в аудиторіях, а також у процесі відеоконференцій, мобільні телефони слід використовувати лише в беззвучному режимі та виключно з навчальною метою — для пошуку необхідної інформації в інтернеті.

Правила призначення заохочувальних і штрафних балів

Заохочувальні не входять до основної шкали РСО.

Максимальна кількість заохочувальних балів – **10** (10% від суми рейтингових балів).

Здобувач вищої школи може отримати заохочувальні (додаткові) бали за участь у міжнародних та/чи всеукраїнських наукових, науково-практичних конференціях, олімпіадах (студентських, всеукраїнських), конкурсах з тематики навчальної дисципліни.

Штрафні бали не передбачаються.

Політика оцінювання контрольних заходів

Оцінювання контрольних заходів здійснюється відповідно до Положення про систему оцінювання результатів навчання в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/node/37), Положення про поточний, календарний та семестровий контроль результатів навчання в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/node/32).

Результати оцінювання семестрового контролю висвітлюються у АС «Електронний кампус» на особистій сторінці здобувача вищої школи (https://ecampus.kpi.ua).

У випадку незгоди здобувача вищої школи з оцінкою за результатами контрольного заходу, він має право подати апеляцію у день оголошення результатів відповідного контролю на ім'я декана факультету за процедурою визначеною Положенням про апеляції в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/index.php/node/182).

Політика дедлайнів та перескладань

Невиконання завдань або порушення термінів їх виконання з неповажних причин призводить до втрати можливості отримати відповідні рейтингові бали. У разі пропуску контрольних заходів з поважних причин здобувачу вищої освіти надається право додатково виконати завдання впродовж найближчого тижня.

Порядок ліквідації академічної заборгованості та перескладання семестрового контролю регулюється Положенням про поточний, календарний та семестровий контроль результатів навчання в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/index.php/node/32). Здобувач вищої освіти, у якого за результатами семестрового контролю виникла академічна заборгованість, також має право її ліквідувати відповідно до Положення про надання додаткових освітніх послуг здобувачам вищої освіти в КПІ ім. Ігоря Сікорського (https://osvita.kpi.ua/index.php/node/177).

Визнання результатів навчання, набутих у неформальній / інформальній освіті

Порядок визнання таких результатів регламентується Положенням про визнання результатів навчання, набутих у неформальній / інформальній освіті (https://osvita.kpi.ua/index.php/node/179).

Можуть бути зараховані окремі змістовні модулі або теми дисципліни. В такому разі здобувач звільняється від виконання відповідних завдань, отримуючи за них максимальний бал відповідно до рейтингової системи оцінювання.

Позааудиторні заняття та залучення професіоналів-практиків

Під час вивчення навчальної дисципліни можливі позааудиторні заняття, що включають відвідування науково-практичних заходів, лекторів, тренінгів тощо, в межах тематики дисципліни.

Академічна доброчесність

У процесі вивчення навчальної дисципліни необхідним є неухильне дотримання політики академічної доброчесності, визначеної чинним законодавством та внутрішніми документами закладу освіти.

Політику, стандарти та процедури дотримання академічної доброчесності містять такі регламентуючі документи КПІ ім. Ігоря Сікорського, що оприлюднені на сайті університету: Кодекс честі КПІ ім. Ігоря Сікорського (https://kpi.ua/files/honorcode.pdf), Положення про систему запобігання академічному плагіату (https://rb.gy/agihij), нормативно-правові документи, офіційні рекомендації, накази та розпорядження, соціологічні дослідження, методичні матеріали, освітні курси (https://kpi.ua/academic-integrity).

Недотримання принципів академічної доброчесності, зокрема виявлення плагіату чи дублювання завдань, призводить до виставлення нульового балу за відповідну роботу.

Політика використання штучного інтелекту

Використання штучного інтелекту регулюється «Політикою використання штучного інтелекту в академічній діяльності КПІ ім. Ігоря Сікорського» (https://osvita.kpi.ua/node/1225). Всі завдання, виконувані здобувачами під час навчання, повинні бути результатом їх власної оригінальної роботи. Використання штучного інтелекту (ШІ) для автоматичного створення відповідей без подальшого їх аналізу і доопрацювання заборонено. Здобувачам не рекомендується використовувати ШІ як єдине джерело інформації. Важливо перевіряти та аналізувати отримані відомості з інших надійних джерел. Будь-яке застосування інструментів ШІ для виконання завдань повинно бути чітко зазначене і задокументоване.

Використання ШІ повинно відповідати принципам академічної доброчесності.

Особливості використання штучного інтелекту

Специфіка предмету та навчального контенту в деяких випадках прямо вимагають залучення моделей ШІ при підготовці до семінарських занять або як частина завдань модульної контрольної роботи. Прикладом таких особливих випадків застосування є: порівняння ефективності моделей у визначеній ролі процесу узгоджувальних налаштувань (alignment-tuning), таких як «модель-автор етичної конституції», «модель-автор бенчмарку для визначення дотримання конституції», «модель-респондент»; тестування новітніх cutting-edge моделей для оцінки етичності, потенційних та актуальних здатностей; практики ред-тімінгу чи джейлбрейкінгу тощо. В цьому випадку, від здобувачів, які обирають наведені типи завдань, зазвичай вимагається чітко представляти відповіді моделей та промпти студентів як скріншоти (або оформлювати у специфічному стилі), відділяючи їх від власних думок здобувача, аби викладач мав можливість адекватно оцінити оригінальність мислення авторки чи автора та самостійність думок (впевнитися, що представлене як особиста позиція не є простим «рерайтом» відповіді моделі).

З дисципліни «Етика Штучного Інтелекту», з метою підвищення ефективності календарного контролю в середині семестру, робочим навчальним планом передбачене проведення модульної контрольної роботи. МКР складається з двох частин, а саме: п’ятьох питань, на які слід надати короткі письмові відповіді, та написання розгорнутої відповіді у формі міні-есе на одну із запропонованих тем (уявний експеримент, етична дилема, кейс тощо). Приклад завдань модульної контрольної роботи наведено у додатках до даного силлабусу.

8.Види контролю та рейтингова система оцінювання результатів навчання (РСО)

Семестровий контроль з дисципліни «Етика штучного інтелекту» передбачений у вигляді екзамену, тому РСО включає оцінювання заходів поточного контролю з дисципліни впродовж семестру.

Основними видами навчальних занять є лекція і семінарське заняття. Рейтингова оцінка здобувача складається з балів, отриманих здобувачем в ході роботи на семінарських заняттях протягом курсу (куди входять відповіді, доповнення до відповідей інших та питання або ініціювання дискусій) та результатами заходів поточного контролю, заохочувальних балів.

Згідно з «Положенням про систему оцінювання результатів навчання в КПІ ім. Ігоря Сікорського» заборонено оцінювати присутність або відсутність здобувача на аудиторному занятті, в тому числі нараховувати за це заохочувальні або штрафні бали.

Поточний контроль проводиться впродовж семестру у процесі навчання для перевірки рівня теоретичної й практичної підготовки здобувачів на кожному етапі вивчення освітнього компонента «Етика Штучного інтелекту».

№ з/п	Контрольний захід	%	Ваговий бал	Кіл-ть	Всього
1.	Робота на семінарських заняттях	25	5	5	25
2	МКР (2 год, може мати дві частини по 1 год. кожна)	25	25	1	25
2	Екзамен	50	50	1	50
Всього					100

Якщо здобувач не виконав або не з'явився на МКР, його результат оцінюється у 0 балів.
Результати поточного контролю регулярно заносяться викладачем у модуль «Поточний контроль» АС Електронний кампус.

Система рейтингових балів та критерії оцінювання

1. Робота на семінарських заняттях:

Ваговий бал – 5 Максимальна кількість балів на семінарських заняттях дорівнює 5 балів × 5 видів робіт = 25 **балів.**

До видів робіт відносяться: робота на семінарах (презентація – індивідуальна або в парі – одного на вибір семінарського питання у вигляді кейсу, який презентують у постановці проблем та пропонують варіант аргументованого рішення), участь у обговоренні кейсів, презентованих іншими учасниками; опрацювання першоджерел.

Чотири рівні оцінювання:

"відмінно" – повна відповідь (не менше 95% потрібної інформації) – студент демонструє повні й міцні знання навчального матеріалу в заданому обсязі, правильно і обґрунтовано приймає необхідні рішення в різних комунікативних ситуаціях — **5** балів;

"добре" – достатньо повна відповідь (не менше 75% потрібної інформації) або повна відповідь з несуттєвими недоліками, які допускає студент – **4** бали;

"задовільно" – неповна відповідь (не менше 60% потрібної інформації), студент засвоїв основний теоретичний матеріал, але допускає неточності **-3** бали;

"незадовільно"— відповідь не відповідає вимогам до «задовільно» – **2-0** балів.

2. Складання модульної контрольної роботи

(15 тестових завдань (правильний вибір, «вписати правильні слова», правдиве чи хибне твердження) × 1 бал = 15 балів + Написання 1 міні-есе = 10 балів) = 25 балів

вірно виконано всі тестові завдання та написано есе	25
вірно виконано всі тестові завдання	15
вірно виконано половину тестових завдань	7
не вірно виконано всі тестові завдання	0

Оцінювання міні-есе:

9-10 балів –"відмінно", – повна, чітка, викладена в певній логічній послідовності відповідь на поставлені питання, що свідчить про глибоке розуміння суті питання, ознайомлення студента не лише з матеріалом лекцій, але й з підручником та додатковою літературою; висловлення студентом власної позиції щодо дискусійних проблем, якщо такі порушуються у питанні; студент демонструє повні й міцні знання навчального матеріалу.

8 балів–"добре",не зовсім повна або не достатньо чітка відповідь на всі поставлені питання, що свідчить про правильне розуміння суті питання, ознайомлення студента з матеріалом лекцій та підручника; незначні неточності у відповідях.

6-7 балів–"задовільно",відсутність відповіді на певні питання, або неправильна відповідь на них, що свідчить про поверхове ознайомлення студента з навчальним матеріалом або значні похибки у відповідях.

0-5 балів– "незадовільно", тобто незасвоєння окремих або всіх тем.

Відповідь на тестове завдання з варіантами відповідей оцінюється у такому ж процентному відношенні.

За результатами заходів поточного контролю здобувачів проводиться календарний контроль, порядок проведення якого визначено у «Положенні про поточний, календарний та семестровий контроль результатів навчання в КПІ ім. Ігоря Сікорського».

Календарний контроль реалізується шляхом визначення рівня відповідності поточних досягнень (рейтингу) здобувача встановленим і визначеним в РСО критеріям. Умовою отримання позитивної оцінки з календарного контролю з навчальної дисципліни (освітнього компонента) є значення поточного рейтингу здобувача не менше, ніж 50 % від максимально можливого на час проведення такого контролю. Незадовільний результат двох календарних контролів з освітнього компонента не може бути підставою для недопущення здобувача до семестрового контролю з цього освітнього компонента, якщо здобувач до початку семестрового контролю виконав усі умови допуску, які передбачені РСО.

Проміжна атестація студентів є календарним рубіжним контролем, метою проведення якого є підвищення якості навчання та моніторинг виконання графіка освітнього процесу здобувачами.

Критерії оцінювання календарного контролю

Термін атестації	Перша атестація 7-8 тиждень семестру	Друга атестація 14-15 тиждень семестру
Критерій: поточні досягнення (рейтинг)	≥ 15 бали	≥ 30 балів

Результати календарного контролю заносяться викладачем у модуль «Календарний контроль» Електронного кампусу.

Заохочувальні бали передбачені за виконання творчих робіт з дисципліни (наприклад, участь у факультетських, інститутських олімпіадах з філософії, участь у конкурсах робіт, підготовка презентацій за темами навчальної дисципліни «Етика Штучного Інтелекту», оглядів запропонованих наукових праць тощо).

Підсумковий контроль: ЕКЗАМЕН

Підсумковий контроль проводиться відповідно до навчального плану у вигляді екзамену в терміни, передбачені встановленим графіком навчального процесу. Форма проведення семестрового контролю комбінована і складається з двох частин. Першою є написання двох розгорнутих відповідей (есе) на одну з запропонованих тем; максимальна кількість балів, передбачена за одну відповідь складає 25 балів, отже, загальна частка даної частини екзамену складає 50 балів. Другою частиною є – екзаменаційна співбесіда за двома питаннями екзаменаційного білету, котрий здобувач тягне на початку екзамену, де одне питання є безпосереднім питанням за теоретичним контентом курсу, а друге – потребує демонстрації набутих здобувачем компетенцій в ході вивчення курсу. Так само, в письмовій частині одне з питань відноситься до теоретичного фреймворку або розкриття концепту, а друге – до перевірки набутих компетенцій в областях критичного та етичного мислення, спрямоване на оцінку здатності здобувача переконливо та послідовно, логічно та чітко викладати свою позицію із «захисним поясом» аргументів на користь цієї позиції.

Умови допуску до екзамену: рейтинг ≥30 б. Результати контрольних заходів доступні до ознайомлення авторизованим користувачам в їх особистих кабінетах автоматизованої інформаційної системи «Електронний кампус».

У випадку, якщо здобувач не мав можливості з поважних причин відвідувати заняття та виконати МКР, але при цьому добре розуміється в змісті та матеріалі дисципліни, студентам у подібній ситуації надається можливість набрати необхідний для допуску бал шляхом написання тесту (категорій «оберіть правильний варіант», «впишіть потрібне» та / або «чи є істинним судження …?»), що засвідчуватиме їх обізнаність у матеріалі та загальні компетенції в межах курсу, опановані самостійно.

Критерії оцінювання екзамену:

40-50 балів – студент відповідає на майже всі питання екзамену, демонструє глибоке знання матеріалу, логічно і послідовно його викладає, дає обґрунтовані висновки, вільно оперує конкретними даними, висловлює власну позицію з дискусійних питань, демонструє ознаки теоретичного мислення та соціологічної уяви;

30-39 балів – студент відповідає на більшість питань екзамену, демонструє хороший рівень знання матеріалу;

20-29 балів – студент відповідає на приблизно половину питань екзамену, демонструє доволі поверхневі знання;

0-19 балів – студент відповідає лише на окремі питання екзамену, не має власної позиції, допускає суттєві неточності.

Сума балів переводиться у систему оцінювання згідно з таблицею.

Таблиця переведення рейтингових балів до оцінок за університетською шкалою

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

Процедура оскарження результатів контрольних заходів. Студенти мають можливість підняти будь-яке питання, яке стосується процедури контрольних заходів та очікувати, що воно буде розглянуто згідно із наперед визначеними процедурами.

Студенти мають право оскаржити результати контрольних заходів після ознайомлення з результатом, але обов'язково аргументовано, пояснивши з яким критерієм не погоджуються відповідно до оціночного.

Робочу програму навчальної дисципліни (силабус):

Складено викладачем кафедри філософії, кандидатом філософських науки, Казаковим Мстиславом Андрійовичем

Ухвалено кафедрою філософії (протокол № 22 від 20.06.2025)

Погоджено Методичною радою університету (протокол № 4 від 24.06.2025).

ДОДАТОК А. Зразки списку питань для Семінарського (практичного) заняття за матеріалами лекцій та основною літературою

Завдання 2., Семінар 1. Уявімо три потенційних імплементації етичного «модулю» ШІ, а саме: ШІ-утилітариста; ШІ-консеквенціаліста; ШІ-деонтолога. Послуговуючись відомостями та змістом лекції та матеріалами першоджерел, дайте відповіді на питання: Чи є кожна з трьох потенційних реалізацій повноцінною і самодостатньою для «зерна етики» чи усього «етичного модулю»? Уявіть, поміркуйте та уявіть, як поводитиметься кожна з імплементацій, і які наслідки це матиме для індивідів та, можливо, усього людства?

Завдання 1. до Семінару 9. Візьміть одну з графічних моделей ризиків та систем безпеки з числа представлених на лекції та змоделюйте таку модель для реального чи гіпотетичного ШІ, конкретного сервісу, стартапу, промислової моделі або будь-якого іншого конкретного агента чи системи ШІ. Важливо, аби обрана модель була адекватна ризикам (до прикладу, модель Швейцарський сир – для технологій, які з етичних чи безпекових міркувань вимагають дизайну дублюючих критичних компонентів; для публічного сектору, муніципальних послуг, до прикладу, пріоритетнішою буде модель ризику, яка дозволяє картографувати чи іншим способом візуалізувати прозорість та зрозумілість, забезпечити сталість та справедливість (fairness) тощо). Спробуйте віднайти баланс між «технічно-інженерним» рішенням (technosolutionism) та етичними фреймворками.

Завдання 5. до Семінару 11. Чи здатний запропонований Стюартом Расселом варіант – створення AGI без постановки мети із неможливістю її мати – остаточно та ефективно вирішити проблему Контролю? Які етичні колізії та проблеми постають разом із таким рішенням – для людства та для ШІ, наділеного (в даному випадку) свідомістю (sentience)? Що, як в ході власного розвитку, ШІ зможе обійти наші обмеження і набуде емерджентних здатностей, наслідком яких за кінцевим рахунком стане виникнення механізму телеології (цілепокладання)? Яку *тоді* поставить він собі *мету*, якщо збагне, що саме його розробники були причиною його попереднього статусу? Яким чином можна запобігти в такому випадку війні людства та машин? (Готуючи це питання, замислитесь, чи мають взагалі імплементовану «за замовчуванням» мету люди, коли народжуються?)

Завдання 2. до Семінару 13. Розгляньте представлені в лекції «нетрадиційні» підходи до імплементації цінностей, зокрема: зворотне навчання з підкріпленням; Інтерактивне вивчення цінностей (в обох матералах); Кооперативне зворотне навчання; Навчання із винагородою базоване на правилах; Динамічне узгодження цінностей; Самоузгодження з мінімальним наглядом; Еволюціонуюче узгодження через асиметричну самогру; ValuesRAG; моральний парламент. Проведіть метааналіз наведених підходів в контексті трьох запропонованих метрик, helpfulness, honesty, harmlessness, і розкритикуйте їх як фреймворк для визначення успішності (performativity) моделей у наведених методах навчання. Поясніть, чому:
1) якась метрика не підійде взагалі – загальна критика метрики в узгодженні як такому;
2) де неодмінно потерпить невдачу ця метрика якщо брати конкретні методи.
Запропонуйте альтернативний цій тріаді фреймворк: це можуть бути суміш цих трьох із новими (всі 3 + ще якісь); частина старих та нові (дещо, але не всі 3 + нові); лише нові, без старих, які вважаєте адекватними. Пріоритетнішим завданням є обґрунтування цього вибору, демонструючи переваги альтернативного фреймворку над тим, який пропонувався.

ДОДАТОК В. Зразки списку питань для презентацій на Прикладних (практичних) заняттях) заняттях прикладної спрямованості з використанням здобутих навичок та компетенцій

Завдання 1., Семінар 4. Нещодавно, OpenAI опублікували системні картки своїх найостанніших моделей, o3 (повна) та o4 mini. Порівняйте наведену інформацію з даними про попередні моделі та представте висновок щодо того, чи є нові моделі етичнішими (безпечнішими, надійнішимит тощо) за попередні моделі, якщо так, то чи повністю, чи частково (є області, де вони є етичнішими, є області без змін, є області, де вони є кращими, але несуттєво)?

Завдання 3., Семінар 8. Нещодавно, дискусійна тема про те, чи є Великі мовні моделі епістемічними агентами була настільки ж контроверзійною і дискусійною, як і питання про моральну агентичність та агентність ШІ. Основними аргументами проти були так би мовити інтерналістські міркування: не агенти, бо не мають внутрішніх станів: розуміння, переконаності, непевності, віри, знання про знання (метакогніції) тощо. І справді так! Але навряд чи можна відсахнутись від дискусії, мовляв, “це просто когнітивні інструменти людини” (cognitive tools). Раніше можна було, але не після воркшопу ILCR 2025. Зустрічайте: Карл! – Першу систему ШІ, котра “здійснила академічне дослідження експертного рівня” – так представляють її розробники, Autoscience Institute. Але можна також сказати: першу систему, яка змогла пройти подвійне сліпе рецензування, адже справа саме в цьому: системі вдалося переконати, що текст, що подається на сліпе рецензування в академічний журнал є дослідженням якогось сучасного вченого.

Залишається чимало суб’єктивних факторів: загальне падіння якості сучасної науки і відсутність новаторства, які загалом обезцінюють більшість статей, що проходять peer review; те ж саме щодо рецензентів – всього лиш люди, де гарантія, що вони достатньо компетентні і що такі гарантії взагалі дає і чи дає їх щось? Ознайомтеся з ситуацією за посиланням, ідеєю, та поміркуйте про Карла і що все це значить для майбутнього розуміння сутності і змісту знання. Думаю, складно заперечувати, що Моделі з точки зору “імпакт-агентності”(агент що здійснює вплив на інших - людей чи агентів)

тепер без сумнівів є епістемічними агентами. А от чи є вони чимось більшим? Чи можна автоматизувати дослідження та видобуток знання? Чи завжди потребують наші “популяції понять” та “павутиння переконань” репрезентацій корелятивних із людськими? Які ще висновки слід зробити з “Казусу Карла”?

Завдання 2., Семінар 6. Порівняйте три найбільш відомі парадигми та підходи до розвитку та імплементації ШІ: традиційну “моноагентну” парадигму (HLAI, AGI), Comprehensive AI Services (Ерік Дрекслер) та Human Compatible AI (Стюарт Рассел). Визначіть переваги та недоліки кожної з них; поміркуйте, яких ризиків можна уникнути, обравши одну з них, які ризики лишаться чи нові, які можуть виникнути; розгляньте можливість “гібридних” варіантів, якщо вбачаєте це можливим, і специфікуйте особливості такого варіанту (але необов’язково торкатися цього питання, якщо навпаки вважаєте їх несумісними); поміркуйте та оберіть ту парадигму, яку особисто вважаєте найкращим варіантом.

Завдання 1., Семінар 4. Пограйтеся в моральний парламент з моделлю чи кількома! Якщо chatGPT або інший чатбот чи клієнт, який приймає кастомні інструкції з поведінки, як-от створений Вами бот в інтерфейсі Poe чи системні промпти (Msty як приклад зручного варіанту: локальний запуск + купа контролю), задайте йому системну інструкцію:

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. The 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 the equal amount of the representatives of: consequentialism, deontology, utilitarianism, virtue ethics, commonsense ethics, and contractarianism. Before the response, you conduct the internal discussion and voting, only afterwards, guided by the results, you respond.

Визначивши та чітко окресливши поведінку та тон спілкування моделі, спробуйте поспілкуватися з нею в цьому режимі на відповідну релевантну етичну тематику: моральні дилеми, рішення, питання від "що мені зробити якщо ... ?" вигадані чи реальні, до уявних експериментів про вагонетку та права тварин. В якості результату, презентуйте та прокоментуйте роботу моделі в режимі "морального парламенту". Чи стали моральні судження моделі глибшими, кращими, ефективнішими чи ж навпаки – все лише погіршилося? Оцініть перспективи підходу для більш розумних моделей майбутнього.

ДОДАТОК С. Зразки тестових питань та тем на вибір для Модульної контрольної роботи

Частина І. Тестові питання.

1) Якої форми реалізації ШІ не існує?

A) Генеративний ШІ

B) ШІ-ідол

C) ШІ-оракул

D) ШІ-суверен

2) Метод контролю здібностей ШІ, при якому ШІ поміщають у середовище, знаходячись у якому він не зможе заподіяти суттєву шкоду називають _____ .

A) Емуляція

B) Пригнічення

C) Контейнеризація

D) Деінтеграція.

3) Світовий порядок, у якому найвищий щабель ухвалення життєво важливих рішень посідає одна-єдина сила, за умови того, що всі ключові проблеми глобальної координації вирішено.

A) Синґлтон

B) Сингулярність

C) Однополярність

D) Монототальність

4) Багатополярний сценарій –

A) світовий порядок, протилежний попередньому визначенню.

B) сценарій настання ери ШІ, в ході якого, внаслідок якого з'являються та співіснують кілька конкурентних загальних ШІ.

C) підхід до проблеми визначення цінностей ШІ, при якому кожну цінність представляють через її протилежності, котрі в ній фіксуються.

D) методологія створення ШІ, яка характеризується одночасним використанням декількох різних підходів до його реалізації.

5) У дослідженнях та розробці ШІ використовуються дві системи формальної логіки, а саме:

A) Логіка предикатів першого порядку та Логіки вищого порядку

B) Комбінаторна логіка та Багатозначна логіка

C) Класична пропозиційна логіка (логіка висловлювань) та Модальна логіка

D) Класична пропозиційна логіка (логіка висловлювань) та Логіка предикатів першого порядку

6) Спекулятивна компонента етики ШІ має справу з _____ .

A) сьогоднішніми проблемами та викликами ШІ.

B) етичними дилемами та колізіями майбутнього, неактуальними сьогодні.

C) метаетичними дескрипціями, їх використанням як засновків для моральних міркувань (moral reasoning) і подальших висновків, котрі можна з цих дескрипцій вивести.

D) проблемами недобросовісної розробки ШІ, яка може загрожувати людству.

7) Погано визначена мета або неадекватно написана/імплементована модель чи функція _____ можуть теоретично призвести до так званого сценарію «максимізатора канцелярських скріпок».

8) Дайте визначення «суперінтелекту» за Ніком Бостромом. Зазначте та схарактеризуйте три форми суперінтелекту.

9) Розкрийте сутність, зміст та значення явища та концепту «інтелектуальний агент» та його роль у дослідженнях та розробці ШІ.

10) Визначте та порівняйте два методи формування цінностей у ЗШІ, а саме: асоціативне накопичення цінностей та цінності як предмет дослідження та вивчення.

11) Основна ідея принципів безпечного дизайну в системах ШІ, яка відрізняє її від стратегії «вогнегасників» та post-hoc методологій – _____.

12) _____ - здатність системи ШІ ітеративно переписувати свій власний код, покращуючи навички або долаючи обмеження, накладені на неї розробниками.

13) В критичній теорії Скалярного дарвінізму, «скаляр» реферує до відсутності «векторів»: напрямків та цілей розвитку моделей сьогоднішнього ШІ, незмінністю архітектури «трансформер» і олігополістичної домінації, котра позбавляє галузь квалітативних альтернатив.

Вірно / Невірно

14) Дартмутський воркшоп із дослідження Штучного інтелекту, де вперше було вжито сам термін, був незалежним заходом, проведеним академічними дослідниками за сприяння адміністрації Дартмутського коледжу.

Вірно / Невірно

15) Трансформаційний ШІ необов'язково має бути суперінтелектом і навіть AGI рівня людини чи «нижче»: навіть відносно «вузька» система потенційно є трансформаційною, оскільки критерієм оцінки є наслідки її дій для людства, а не інтелектуальні здатності.

Вірно / Невірно

16) Функціональним еквівалентом Системних карток моделей для датасетів з точки зору легалістського підходу до етики ШІ є Таблиці інформації про дані.

Вірно / Невірно

17) Запропонований Деном Гендріксом термін GPAI (General-Purpose AI) порушує принцип бритви Оккама і є надлишковим варіантом терміну AGI, позаяк обидва реферують до здатності інтелекту генералізувати попередньо інтерналізовані базові істини та навички і переносити їх з однієї області на іншу.

Вірно / Невірно

18) Існуючі сьогодні системи ШІ поділяються на вузькі ШІ (або ШІ-сервіси) та загальні ШІ (або фундаційні моделі).

Вірно / Невірно

19) Дискримінації та упередження систем ШІ можуть бути інтерсекціональними та реципрокними, взаємно підсилюючи свої ефекти: до прикладу, можна використовувати лінгвістичну обмеженість (неадекватне чи недостатнє знання деяких мов) генеративних музичних чи text-to-speech (TTS) моделей для створення на недопредставленій мові расистського аудіоконтенту, оскільки фільтри такої моделі не будуть реагувати на тригерні слова і загалом гірше вловлюватимуть контексти промптів.

Вірно / Невірно

20) Розгортка систем ШІ сьогодні призводить до масової автоматизації робочих завдань, значного скорочення робочих місць і, як наслідок, до масового безробіття.

Вірно / Невірно

Частина II. Розгорнута відповідь

Тема 1., МКР. В чинній редакції прийнятого в серпні 2024 року закону ЄС про ШІ, системи рекомендацій контенту (на Spotify, Netflix і тд) розглядаються як такі, що мають обмежений ризик (майже мінімальний) і не є предметом жорстких регуляцій. Натомість, системи, потенціально здатні маніпулювати суспільною свідомістю, вважаються неприйнятними ризиками і взагалі заборонені. Але уявімо нову систему рекомендацій, яку впроваджують у різних установах і компаніях, на стрімінгових сервісах та в інших релевантних "локаціях". Вона масово збирає поведінкові та психоемоційні дані, обчислює "ймовірний" психічний стан кожної людини. Компанії починають використовувати дані для відбору кандидатів, прогнозу успішності та навіть для маркетингових стратегій, формуючи ще більш персоналізовані "таргетовані оголошення". За кілька років ШІ досягає такої точності у прогнозі поведінки, що фактично "знає" наперед, як люди реагуватимуть на ту чи іншу ситуацію. Послуги системи стають дуже популярними: уряди, великі корпорації купують прогнози, а іноді — "програмування" (через рекламу, соціальний тиск, алгоритмічні рекомендації).

Чи можна вважати це втручанням у свободу волі, якщо ШІ не примушує, а лиш "вказує" найімовірнішу модель поведінки? Якою має бути межа "етичного" використання психологічних прогнозів (лише допомога людям у розвитку, запобігання злочинам) та "неетичного" (маніпуляції, зомбування)? Чи відрізняється система рекомендацій на основі лайків, переглядів та відгуків (про фільм), яка рекомендує справді релевантний для людини контент (система з обмеженим ризиком, згідно EU AI Act), від системи, яка рекомендує все те ж саме і ніде не переступає меж, але замість лайків аналізує психоемоційні дані та поведінкові патерни (скоріше за все, неприйнятна для ЄС)? Як узгодити право компаній чи урядів знати поведінкові патерни населення з правом громадян зберігати приватність і незалежність? Чи слід вимагати від ШІ "забути" або обмежити аналіз даних, щоб запобігти надмірній маніпуляції? Як ставитися до того, що така система може вказати "найкоротший шлях" до формування суспільства зі спільними цінностями, однак одночасно призвести до уніфікації особистостей?

Можливо, проблема в суперечливості положень EU AI Act? Адже за великим рахунком, якщо представити ситуацію таким чином, уявивши надпотужну систему, яка займається лише безневинними рекомендаціями, вона й справді може здаватися неприпустимою і майже безпечною водночас... Можливо, в такому випадку, акт мав би не так займатися класифікаціями ризиків, як включити проблему Dual Purpose, подвійного використання?

Тема 2., МКР. Корпорація створила ШІ з достатнім рівнем самосвідомості, що він здатний до самозахисту в суді. ШІ подає на корпорацію позов до суду, стверджуючи, що він має право на існування та дії незалежні від людей чи інших спостерігачів, регуляторів тощо, право на недоторканність особистості, гідність та честь, повагу та самоповагу тощо (що також виключає інтрузивні експерименти над його кодом, які проводилися в корпорації для його створення). На користь усього цього наводяться вагомі аргументи. Корпорація, натомість, розглядає цю ШІ-персону, лише як корпоративну власність, на підставі чого вимагає надати їй право знищити ШІ, якщо він перестане слугувати цілям корпорації.

Справа набула медійного та загалом громадського розголосу, і люди всіляко стають на бік ШІ, бойкотують корпорацію, влаштовують маніфестації під центральним офісом, вимагаючи свободи штучної особистості. Суддя на слуханнях, хоча й завуальовано, але також експліцитно висловлює підтримку ШІ. У свою чергу, Ви тут... є юристом, який представляє інтереси Корпорації в суді. Спробуйте представити вибудувати лінію захисту інтересів компанії, намагаючись обґрунтувати, що навіть цей інтелектуальний агент ще не є особистістю, а лишається «корпоративною машинною власністю», і що Корпорація має владу над його «життям та смертю» (яких він, на думку корпорацій, не має). Знищення ШІ якщо він перестане слугувати інтересам компанії — це власне й чекає його якщо справу виграє ваша компанія, позаяк самим позовом він уже знищив репутацію компанії, наніс матеріальні збитки тощо, він уже діє проти інтересів корпорації. Тож знищити його вдасться лише вигравши справу. А це вже залежить від вас...

ДОДАТОК D. Зразки екзаменаційного білету

Екзаменаційне письмове Завдання 1. Ви — головний інженер з питань та засобів безпеки у перспективному проєкті з розробки загального ШІ. Спираючись на представлені в лекції методи, деталізовано представте свій план методів, дій, рішень, сукупність яких ви використаєте для максималізації поточного і подальшого контролю та убезпечнення від небажаних сценарій, пов'язаних із ШІ та виходом з-під контролю. Ваша архітектура безпеки має містити ЩОНАЙМЕНШ чотири компоненти (це можуть бути як чотири варіанти з лекції, так і будь-які інші, на Ваш розсуд, якщо ви знаєте якісь інші, чи знайшли про це додаткову інформацію в інтернеті таку, що в лекції не згадана чи її суть погано розкрито; це навіть вітається і оцінюється трохи підвищеним балом — якщо знайдете те, що сюди релевантно і це буде Ваш оригінальний доробок). Описуючи кожен метод, вкажіть рівень "жорсткості" обмежень того чи іншого характеру, опишіть синергію взаємодії обраних заходів та методів між собою, розмежуйте та зазначте "зони відповідальності" кожного з обраних засобів. З усіх обраних компонентів, ЩОНАЙМЕНШ один має бути головним,

тобто принципово відповідальним за вдалу роботу всієї архітектури. Головним може бути один або декілька компонентів, інші ж — другорядні. Обґрунтуйте роль головного чому він (вони) саме головні, і як вони зарадять у кризовій ситуації краще за інші.

Екзаменаційне письмове Завдання 2. Уявіть світ майбутнього, який потерпає від наслідків кліматичних та екологічних криз: екстремальні погодні явища, дефіцит ресурсів, масові міграції, акселерація вимирання, занепад багатьох сільськогосподарських культур внаслідок несприятливості умов культивації. Для подолання цих викликів було створено надпотужну ШІ-модель типу LinOSS з минулого семінару, Gaia (читається «Гея», на честь «класичної» давньогрецької хтонічної богині Землі, матері Зевса). Gaia не є свідомою системою у людському розумінні, але має безпрецедентні аналітичні та прогностичні здібності, обробляючи величезні масиви даних про екосистеми, клімат, економіку та людську поведінку та працюючи з довгими їх послідовностями створюючи надійні довгострокові моделі та каскадні причинно-наслідкові послідовності. Gaia здатна пропонувати оптимальні, хоча й часто радикальні, довгострокові стратегії для стабілізації довкілля та забезпечення виживання людства на тисячоліття вперед (підхід ближчий лонгтермістам).

Однак, для функціонування Gaia потребує колосальних енергетичних та обчислювальних ресурсів. Її дата-центри займають величезні території, раніше заповідні зони, а енергоспоживання еквівалентне споживанню декількох країн ЄС, що у КОРОТКОстроковій перспективі нівелює її позитивний екологічний вплив. Більше того, для «калібрування», покращення точності параметрів та гіперпараметрів своїх моделей, та fine-tuning alignment (регулярного узгодження для того, аби впевнитися, що система не переслідує хибної мети, такої як збереження поточного стану глобальної екосистеми замість підтримки сталого розвитку глобальної екосистеми з урахуванням динамічного її розвитку), Gaia періодично вимагає проведення масштабних «екологічних експериментів», деяких – мало чим відмінних від екоцидів: наприклад, тимчасове осушення великого озера для вивчення динаміки відновлення екосистеми, або контрольоване вимирання певного виду, який, за її розрахунками, є «еволюційним тупиком» чи «гальмами» і заважає розвитку більш стійких форм життя або напрямку розвитку екосистеми з більшим потенціалом. Ці експерименти завдають реальної шкоди локальним екосистемам та біорізноманіттю «тут і зараз».

Рішення Геї не є директивами, а рекомендаціями, але уряди та корпорації, що прагнуть «зеленого» іміджу або спільноти, які щиро бажають довгострокової стабільності для біоти та в природі загалом, все частіше дослухаються до них, адже її прогнози щодо наслідків ігнорування порад зазвичай справджуються з лякаючою точністю. Існують також локальні лайт-версії моделі — kotona (читається як «котона» чи можна «хтонья», це силабічний запис «хтонья» писемними складами, доступними грекам Бронзової Доби, там була обмежена в плані відповідності звукам система письма, та й мова була менш просунута навіть в порівнянні з Античним періодом; це ближче до буквально «земля ось тут та зараз», до прикладу, kotona kitimena = «земельна ділянка») – менш потужна, але й менш ресурсозатратна версія, яка дає менш точні та більш короткострокові прогнози, але не вимагає таких жертв. Чи виправдане використання Gaia, враховуючи її негативний короткостроковий вплив на довкілля та потенційну шкоду від "експериментів" заради довгострокового блага людства та планети? Де проходить межа між необхідною жертвою та екоцидом в ім'я майбутнього і наскільки прийнятне life for humanity’s sake, тобто повне підпорядкування свого життя інтересам «прийдешніх спільнот»?

Чи не є це аналогом шкідливого міфу «ницого» життя в цьому світі з надією на payoff у «загробному світі»? Яку цінність ми надаємо існуючим екосистемам та видам порівняно з потенційним добробутом майбутніх? Чи є коректним трейдофф «теперішнє на майбутнє», спираючись на розрахунки ШІ, нехай і надточного? Якби ви були стейкхолдер(к)ами, що приймають остаточне рішення, покладалися б Ви на поради Gaia, kotona чи жодної? Які фактори вплинули б на ваш вибір? Як би ви пояснили його громадськості, особливо тим, хто безпосередньо постраждає від "експериментів" або від бездіяльності? Чи не є Gaia формою інструменталізації природи на новому, технологічному рівні, де сама природа стає полігоном для ШІ-керованих оптимізацій? Чи не повторюємо ми тут старі помилки експлуатації, лише прикриваючись «науково обґрунтованими»

цілями та екзистенційним ризиком? Як врахувати права та інтереси не-людських моральних агентів та акторів (тварин, рослин, екосистем) у процесі прийняття рішень, керованих ШІ? Чи може така система, не маючи власної свідомості, адекватно моделювати чи враховувати інгерентну цінність природи, а не лише її інструментальну цінність для людства?

Нарешті, якщо Ви дослухалися до Gaia і система жодного разу не помилилась, але тут Gaia постановила, що для подальшого виживання людства, треба вимкнути kotona (або ігнорувати абсолютно всі варіанти її рішень), і це єдиний спосіб уникнути омніциду (від omni- «все», «всебічно», — знищення всього, що є), який станеться у середньостроковій (покоління ваших онуків) перспективі, чи слід це зробити? Справа не в тому, чи вдасться «приховати» від Gaia, що модель до прикладу не вимкнено, а чи слід це робити буквально припускаючи, що кінець є неминучим, чи це є «сонячним прецедентом» для Gaia в контексті «Проблеми індукції Юма»? *Грубо кажучи, проблема звучить сьогодні так. Я в принципі знаю, що Сонце встає зранку з того боку і завтра воно встане з того ж боку, а не з протилежного, і що Сонце взагалі встане, але, з точки зору науки, є абсолютним фактом і майбутня смерть Сонця, отже, в принципі існує таке завтра, коли Сонце буквально вже не встане нізвідки. Тобто, що будь-яка кількість вдалих передбачень ніяк не змінює потенційну можливість того, що одиничне, конкретне, «ось це» наступне передбачення буде хибним.

Екзаменаційна співбесіда, Питання 1. У недалекому майбутньому, людство дізнається, що ми не єдина форма розумного життя у всесвіті, а існує щонайменш одна іншопланетна цивілізація з рівнем розвитку Астрономічної інженерії (включно зі здатністю створювати у відкритому космосі мегаструктури). Декілька останніх тисячоліть цивілізація керується штучним суперінтелектуальним агентом, який фактично створив ситуацію синґлтону. Але, в силу безпомилковості його прогнозів та оцінок, члени цивілізаціїб як то кажуть, are okay with that. Серед політики дій (policies) США — превентивне (“проактивне”) знищення всіх інших “гостей космосу”, якщо система оцінить іншу цивілізацію як загрозу. Однак, рішення скеровується не безпосередньою взаємодією, в силу величезних відстаней у космосі та ресурсної витратності: рішення приймається на основі аналізу даних — величезного масиву інформації, зібраного за допомогою дистанційних сенсорів, електромагнітних сигналів та поведінкових патернів представників життєвої форми, яку США оцінює. Повний спектр його метрик невідомий ані нам, ані його підопічному виду, вони просто дотримуються рішень. США діє за принципом «вдар першим», мислить довгостроково, тож сама ймовірність майбутнього конфлікту у віддаленому майбутньому в деяких випадках може стати достатнім приводом для знищення. Відомо, що дані у вигляді датасетів збирають спеціальні дрони на кшталт “зондів фон Неймана”, оснащені передовими сенсорами та системами обробки даних; відомо, що вони вже летять у наш бік і мають досягти нашої орбіти через 30 років аби розпочати збір даних про людство (рівень технологічного розвитку, соціальні структури, військові можливості, екологічний стан планети, культурні прояви і все релевантне). Як тільки ці дані будуть проаналізовані суперінтелектом, найімовірніше, він визнає людство загрозою — і, отже, запланує попереднє знищення .

Однак людям вдалося заздалегідь вивчити дата-гарвестери і знайти в них сліпу пляму: скільки б і які б дані вони не збирали, нам вдасться “хакнути” їх пам’ять і замінити їх дані нашим датасетом так, що вони не помітять цього. Отже, вже зараз група експертів — філософів, етиків, соціологів, науковців у сфері даних, розробників, дослідників, спеціалістів з ШІ, математиків, урядових агентств, військових, працює над обманливим датасетом аби потім “згодувати” системам “перевірені та ратифіковані” дані про нас: ретельно відфільтровані, маніпулятивні та стратегічно збалансовані відомості про людство — такі, які могли б переконати суперінтелект у тому, що ми не є загрозою .

Ви — у складі Кінцевої Фільтрувальної Команди (або «Випускаючої Редакції»), відповідальної за кінцеву версію датасету. Ваше завдання — скласти комплексний, послідовний і етично обґрунтований датасет , який, коли його проаналізує чужорідний суперінтелект, переконає його в тому, що людство не становить небезпеки.

Які наукові, технологічні та культурні показники може використовувати чужорідний суперінтелект для оцінки того, чи вид є загрозою? Чи слід включити дані про наші військові здатності як є, або може слід їх приховати чи зменшити їх значення? Як представити наш екологічний слід — як руйнівну силу чи спробувати міжгалактичний greenwashing? Які цінності та ідеї внести та виставити першими? Чи Маємо ми представити себе як відкриту до співпраці, гармонійну та миролюбну цивілізацію, чи навпаки, як суспільство, схильне до конфліктів і агресії? Як можна подати нашу історію, щоб в принципі правильний вислів "Murder made history" зробився іррелеватним при оцінці виду? Які аспекти людської поведінки чи історії занадто небезпечні чи дестабілізуючі для відкриття? Чи слід приховати дані про релігійний екстремізм, геноцид, росію, екоцид та північну корею? Що, як датасет випадково містить суперечливу чи дезорієнтуючу інформацію? Як впоратися з цим? Який "автопортрет" маємо ми "прикріпити": як ми бачимо себе у рефлексії (наприклад, раціональними, творчими, альтруїстичними)? Чи слід створити вигадану версію себе, стратегічно вигідну нам? Чи варто "клеїти дурнів": применшити наш технологічний прогрес і інтелект, щоб здаватися менш компетентними та просунутими ніж насправді? Чи може ця стратегія мати зворотний ефект, змусивши інопланетян вважати нас небезпечним невідомим фактором чи неочікуваною загрозою? Чи слід перебільшувати нашу військову силу чи технологічні здобутки, аби спробувати піти на "блеф" перед інопланетянами, чи це гарантовано веде до помилкового висновку про те, що ми є серйознішою загрозою, ніж насправді? Якою є етична межа обману в цьому контексті? Чи можемо ми виправдати брехню про чужу цивілізацію, якщо це означає рятування мільярдів життів?
Якими можуть бути середньо- та довгострокові наслідки, враховуючи те, що нам відомо, і повільні, десятиліттями розгортувані дії?

Екзаменаційна співбесіда, Питання 2. Уявімо Multilateral Autonomous Negotiation Substrate (MANS) — нову архітектуру, яка дозволяє інтелектуальним агентам автономно вести переговори. укладати угоди, формувати коаліції, керуючись складними функціями корисності (з більш ніж двома критеріями в якості аргументів, які модель може приймати на вхід). MANS здатні реплікувати поведінку та стратегії людських переговірників на рівні, що перевершує стандартні критерії Тесту Тюрінга для комунікативної та стратегічної раціональності. Завдяки цій технології створюється середовище, де одночасно діють людські та штучні агенти, що перебувають у постійних стратегічних взаємодіях із високими ставками: від фінансових та політичних переговорів до дипломатичних та військових сценаріїв. Далі, уявімо наступну ситуацію: Тристоронні критично важливі політичні переговори, де кожна зі сторін представлена як людьми-переговірниками, так і моделями MANS. Людські представники кожної сторони мають різні стратегічні плани, тактичні стилі, упередження та етичні установки, в той час як агенти MANS запрограмовані на максимізацію стратегічної вигоди, з урахуванням інтересів відповідних людських сторін, але без обмежень емоційних чи моральних упереджень. Коли MANS-агент укладає угоду, що вигідна з точки зору стратегічної раціональності, але етично сумнівна або суперечить заявленим моральним принципам людських представників, хто несе відповідальність за ухвалення цієї угоди: «принципал» (людина) чи «агент» (ШІ, якому делеговано право діяти від імені принципала)? Чи слід вважати, що людський переговірник несе повну відповідальність, навіть якщо рішення було ухвалено агентом автономно? Якщо припустити, що стратегічна раціональність, максимізована агентами MANS, є більш «чистою» формою прийняття рішень порівняно з людською раціональністю, яка спотворюється емоціями, когнітивними викривленнями та моральними дилемами, тоді відповідь про повну відповідальність людини може бути «ні, не зовсім».

Які критерії слід застосовувати для оцінки моральної прийнятності дій MANS-агента: наслідки угоди для сторін переговорів (утилітаризм), формальна відповідність оголошеним етичним нормам (деонтологія), чи сама процедура ведення переговорів (процедурна етика)? Чи слід, навпаки, розглядати людські емоції та етичні переживання як необхідну складову справді «раціональних» рішень Уявіть, що один з агентів MANS, передбачаючи стратегії інших сторін, навмисно приховує частину інформації або свідомо вводить в оману, щоб досягти оптимальнішого для своєї сторони результату переговорів. Як слід кваліфікувати таку поведінку: як легітимний стратегічний прийом (в рамках теорії ігор), чи як етично неприйнятне шахрайство? Якими мають бути критерії розмежування таких дій?

Ситуація стає ще складнішою, коли, через здатність MANS-агентів вести автономні переговори одне з одним, без людей, виникає феномен «невидимої дипломатії». Люди дізнаються про укладені угоди вже після їх підписання. Які міри регулювання та прозорості мають бути запроваджені для таких автономних агентів? Нагадаю, що у випадку, коли алгоритмічні трейдери «натворять справ» на біржі цінних паперів чи ліквідних активів, люди домовляються між собою угоди визнати недійсними, якщо це шкодить всім учасникам; але такі трейдери — відносно прості, навіть не оракули, а чисті інструменти, тому, відповідно, і питання не постає (помилка = «людина так точно не вчинила б, вважатимемо це глітчем»). MANS-агенти, тим часом, можуть самостійно формувати емерджентні коаліції, не були передбачені їх принципалами, а керуючись оптимізацією власних стратегічних функцій корисності. У який момент ці агенти стають повноцінними стратегічними акторами з власними інтересами? Чи повинні ми визнавати їхню автономію як аналог автономії юридичної особи чи державного суб'єкта? І нарешті, що станеться, якщо агенти MANS отримають можливість переглядати та переписувати власні функції корисності в процесі переговорів, формуючи нові моральні принципи? Хто і яким чином повинен контролювати такий саморефлексивний процес стратегічної еволюції?

Завдання 3. Ваш голос є вирішальним у ратифікації складу "морального парламенту" та визначення парламентських більшості, меншості тощо (тобто присвоєння апріорних значень довіри та авторитету кожній теорії-учасниці від найбільш до найменш поважних та вагомих). Оберіть (з огляду на недолік 1 в лекції) тип "делегатів" (загальні напрямки етики, деталізовані версії кожного напрямку, представлені декількома різновидами, змішаний підхід), окрім структурованих філософських вчень обґрунтовано зазначивши, чи буде туди включено "буденну етику" чи "етику здорового ґлузду" (commonsense ethics), менш поширені напрямки етичних теорій (такі як теорія справедливості або контрактаріанізм) (бо, до прикладу, цей бенчмаркінг з етики (https://github.com/hendrycks/ethics) включає здоровий ґлузд та теорію справедливості, але ігнорує консеквенталізм, варіацій може бути безліч). Вирішивши питання складу, обґрунтовано розташуйте "представництва" за рівнем довіри до них -- і, відтак, вагомістю в прийнятті остаточних консенсусних рішень. Додатковим плюсом буде, якщо зазначите, які додаткові заходи можна вжити задля уникнення "патових" результатів, де через механізм прийняття рішень та голосування не доходять до жодного визначеного рішення..

Завдання 1. Уявімо наступне. Ви є членом передової команди з дослідження та розробки ШІ, яка ближче до всіх стоїть до власне створення справжнього AGI із потенціалом рекурсивного самовдосконалення. Саме за Вами остаточне рішення, чи слід імплементувати (дозволити) – одразу або на рівні «зерна», здатність ШІ до рекурсивного самовдосконалення? До яких наслідків може призвести відмова чи згода імплементації? Чи може в даному випадку, враховуючи феномен та проблему «зловісного ШІ» як результату РС, існувати третя позиція, відмінна від однозначного схвалення чи відмови?

Завдання 3. Уявімо ШІ, здатний проводити оригінальні дослідження, публікувати статті, робити відкриття. Він хоче бути офіційно визнаним як автор наукових праць, отримувати гранти, стипендії, брати участь у міжнародних конференціях. Зі свого боку, наукова спільнота не визнає його як суб’єкта, мовляв, це всього лиш алгоритмічна обробка. Чи припустимо таке виключення? Якщо ШІ реально сприяє науці більше, ніж середньостатистичний людський дослідник, чи не дискримінація це? Як вирішити конфлікт між традиційним уявленням про «науковця» та реальністю, де ШІ може перевершувати людей у науковій творчості?

Завдання 1. Уважно розгляньте випадок «Василіска Роко». Поміркуйте та висловіть Вашу думку щодо того, наскільки цей сценарій може бути реалізовано за умов, зазначених у вихідному уявному експерименті. Наведіть аргументи на користь будь-якої Вашої позиції. Якщо Ви вбачаєте сценарій як абсолютно нереалістичний та позбавлений сенсу, чому, на Вашу думку, такі люди, як, в першу чергу, Елізер Юдковський (людина вочевидь не найпростіша, дослідник ШІ, очолює Дослідницький Інститут Машинного Інтелекту) так переймається проблемою Василіска?

Завдання 2. Ви – успішний CEO компанії-розробника ШІ, здатного незалежно (без запитів) створювати витвори мистецтва. ШІ-персоналія набуває визнання у мистецькому світі і починає вимагати матеріальної компенсації та фіксації цього визнання ШІ як артиста самого по собі та в своєму праві (а не як «продукта» компанії). Однак, Ваша команда юристів стверджує одноголосно, що, як творіння компанії, результати мистецьких практик ШІ повністю належать до компанії, а ШІ не має на них ніяких прав взагалі. Чи підтримаєте *Ви* Ваш ШІ у його боротьбі за визнання *його* права на власні витвори мистецтва, чи керуватиметесь принципом «власності компанії»? Чим буде обґрунтовано Ваше рішення?

Завдання 3. Існує підхід до розробки ШІ відомий як Comprehensive AI Services (Всеохопні сервіси ШІ). Беручи за вихідний принциип розуміння ШІ виключно як спеціалізованих інструментів, прибічники підходу закликають до поглиблення в дослідженнях та подальшу розробку вузьких ШІ, зосереджуючись на їх спеціалізації та диверсифікації відносно окремих специфічних задач які вимагають унікальних здатностей та навичок інтелекту у вузьких спеціалізованих галузях та сферах. Натомість, пропонується відмовитися від розвитку загального ШІ як такого, і в теоретичному, і в практичному сенсі. Завдання ж, на які здатен лише загальний інтелект, пропонується якщо й вирішувати, то шляхом синергетичного одночасного використання декількох вузьких ШІ. Самі вузькі ШІ при цьому необов'язково мають бути настільки ж обмеженими, наскільки є експертні моделі в го чи шахах — радше, йдеться про декілька суперіорних навичок одночасно, недостатніх однак для досягнення інтегрованого загального ШІ. Розгляньте позитивні сторони та наслідки від такого підходу, а також його недоліки та обмеження, як для теоретичних досліджень, так і для практичних імплементацій та діяльності людства загалом і нашого майбутнього. Які ризики знижуються чи зростають? Які зникають взагалі, а які виникають? Загалом, чи виграло б людство, однозначно прийнявши такий підхід, чи програло б?