



ЕТИКА ШТУЧНОГО ІНТЕЛЕКТУ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	Перший (бакалаврський)
Галузь знань	Усі спеціальності
Спеціальність	Усі спеціальності
Освітня програма	Усі спеціальності
Статус дисципліни	Вибіркова
Форма навчання	очна(денна)
Рік підготовки, семестр	2 курс, осінній / весняний семестр
Обсяг дисципліни	2 кредити ЄКТС / 60 год.: Лекційних занять: 16 год. Практичних занять: 14 год. Самостійна робота студентів: 30 год.
Семестровий контроль/ контрольні заходи	Залік, МКР
Розклад занять	http://schedule.kpi.ua/
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: <i>Казаков Мстислав Андрійович</i> <i>Кандидат філософських наук</i> potya@ukr.net Практичні / Семінарські: <i>Казаков Мстислав Андрійович</i> <i>Кандидат філософських наук</i> kazakov.mstyslav@lil.kpi.ua
Розміщення курсу	Платформа дистанційного навчання «Сікорський» (Sikorsky Distance Learning Platform)

Програма навчальної дисципліни

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

Протягом останнього десятиріччя, технології штучного інтелекту (ШІ) прогресують величезними кроками, стаючи ключовою тканиною, що складає нашу реальність. Amazon, котрий персоналізує наш шопінг; GPS-навігація та інші послуги допомагають нам уникати заторів та швидко знаходити оптимальні маршрути; ми покладемося на Google у пошуку необхідної нам інформації; автомобілі стають автономними від водіїв та пасажирів, керуючи собою замість людських індивідів. Синтетичні сутності, як антропоморфні, так і програми, що скеровують нашу взаємодію та комунікацію, навчилися емулювати не лише зовнішню форму, а й внутрішній світ людини, думку, почуття, емоцію і навіть мистецтво. Google Magenta та DALL-E розвиваються у створенні мистецьких творів, згенерованих ШІ; коротке оповідання, написане ШІ, виграв літературний приз у Японії; подібно DeepBlue у шахах, AlphaGo виграв у чемпіона світу з давньої гри Го (китайські шашки), однаково майстерно оперуючи хірургічною точністю логічного мислення та спонтанними кроками, із поєднання яких виковуються переможна стратегія та тактика. ШІ доводить і демонструє фактично, що давнішня мрія людства – досягнути останнього фронтиру створення розумного життя, такого, що може значно перевершити творця в усіх

мислимих сферах інтересів і навичок – більше не далека фантазія чи езотеричні міркування, і не лише звивисті тропи науково-фантастичного генія: це завтрашня реальність, якої не уникнути. А реальність, будь вона фізичною, соціальною, історичною, - не лише даність чи «плюси»; завжди своїм наслідком має вона також проблеми, виклики, загрози. Останнім, в тому, що стосується виміру ІІІ та ІІІ як фрагменту нашої нової – теперішньої та завтрашньої – реальності, і намагається дати раду, запропонувати рішення, нова область міждисциплінарних досліджень, Етика Штучного Інтелекту.

Мета вивчення дисципліни – надати інструментарій та методологію, наукове та техніко-технологічне знання і здатність до критичного осмислення та участі в дебатах в області ІІІ, котра розвивається з невлмовимою швидкістю; метою курсу є дослідження етичних імплікацій в технологіях ІІІ, 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; додатковою метою є культивування у слухачів та слухачок курсу гнучкого та адаптивного мислення при розгляді різного роду проблемних аспектів комп'ютаційної архітектури та шляхів імплементації ІІІ, сприймаючи логічну основу прийняття рішень ІІІ, інкорпоруючи отримані знання та навички у фреймворк соціальної, громадської та етичної відповідальності через широкий спектр варіантів дій, вчинків, рішень, підходів, тактик та стратегій.

Програмні результати навчання.

Набуття здібностей та здатностей до прагматичних імплікацій ІІІ 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 ІІІ загалом;

знання та вміння генерувати нові знання, ідеї, моделі та гіпотези щодо довгострокових та високоабстрактних моральних питань та дилем;

включати імплікації ІІІ до людського сприйняття та особистісності, на рівні буденних та нетривіальних компутаційних інтуїцій щодо оперування ІІІ 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 ІІІ» 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 ІІІ та виходу за межі «вузького» ІІІ до Загального ІІІ та Суперінтелекту – і як екзистенційного ризику, і як потенційного вирішення всіх або частини глобальних проблем людства загалом.

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 ІІІ (машинне навчання, структури даних, нейронні мережі, прикладний ІІІ тощо) – не є обов'язковим; перевагою буде також розуміння ключових концептів соціальних та політичних теорій, таких як справедливість, публічна політика, управління, менеджмент тощо.

Постреквізити: специфічні етичні проблеми (приватність, похибки, агентичність, автономія, управління ІІІ технологіями, регуляція тощо); ІІІ-політика та менеджмент, управлінські фреймворки міжнародного та державного рівнів; дослідницькі методи в області ІІІ, аналіз даних, квантитативні та квалітативні дослідження, міждисциплінарні дослідження.

3. Зміст навчальної дисципліни

Тема 1. Штучний Інтелект, етика та їх інтерсекція.

Тема 2. Штучні моральні агенти та пробіл у відповідальності.

Тема 3. Ризики та виклики штучного інтелекту. Відповідаючи на виклики, знижуючи ризики.

Тема 4. Роблячи ШІ Етичним

Тема 5. Парадигма публічного ШІ як альтернатива пропріетарності

Тема 6. Підходи до створення відповідального та пояснюваного ШІ.

Тема 7. Принципи безпечної розробки та підходи до формування цінностей.

Тема 8. Проблема контролю: дивлячись в теперішнє і майбутнє.

4. Навчальні матеріали та ресурси

Наведено рекомендовані навчальні матеріали та ресурси для засвоєння матеріалу, розглядуваного на лекційних заняттях та практичних заняттях.

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. 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
7. Рассел С. (2020) Сумісний з людиною: Штучний інтелект і проблема контролю. Київ: Форс Україна
8. Сейтуа де, М. (2023) Код творчості. Як штучний інтелект учиться писати, малювати, думати. Київ: Arthuss
9. Шаров С. (2023) Сучасний стан розвитку штучного інтелекту та напрямки його використання. - https://www.researchgate.net/publication/370595289_Sucasnij_stan_rozvitku_stucnogo_intelektu_ta_na_pramki_jogo_vikoristanna

Допоміжна

1. Boddington, P. (n.d.). AI Ethics: A Textbook. Springer Nature Singapore. <https://doi.org/10.1007/978-981-19-9382-4>
2. 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>
3. Coeckelbergh, M. (2020). AI Ethics. The MIT Press.
4. Montemayor, C. (2023). The Prospect of a Humanitarian Artificial Intelligence: Agency and Value Alignment. Bloomsbury Academic Publishing.

Література зі списку, котра може бути відсутньою у вільному доступі для завантаження на момент запиту, буде надана викладачем.

5. Методика опанування навчальної дисципліни (освітнього компонента)

Навчальна дисципліна охоплює 18 годин лекцій та 18 годин практичних занять, а також виконання модульної контрольної роботи (МКР).

Практичні заняття з дисципліни проводяться з метою закріплення теоретичних положень навчальної дисципліни і набуття студентами умінь і досвіду оперувати сучасними поняттями в галузі енергетичного менеджменту. Виходячи з розподілу часу на вивчення дисципліни, рекомендується дев'ять практичних занять (з врахуванням часу на МКР).

Методи та форми навчання включають не лише традиційні університетські лекції та семінарські заняття, 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. Під час навчання та для взаємодії зі студентами використовуються сучасні інформаційно-комунікаційні та мережеві технології для вирішення навчальних завдань.

Лекційні заняття

Лекція 1. Штучний Інтелект. Етика: ключові поняття, історичний огляд, зміст state-of-the-art

1. Штучний Інтелект: Історія ідеї, визначення, поточний стан справ (реалії та спекуляції).
2. Етика: прикладна філософська дисципліна про вчинки, взаємодію та мораль.
3. Етика ШІ як спеціалізована етика, її характеристичні особливості.
- 3.1. Виклики ШІ як виклик етиці.
- 3.2. Етика ШІ як єдина етика зі спекулятивною складовою.
- 3.3. Метаетичні імплікації в питаннях реалізації Загального Штучного Інтелекту.

Лекція 2. Штучні моральні агенти та пробіл у відповідальності.

1. Штучні інтелекти: реальні та можливі
- 1.1. Рівні інтелектуальності
- 1.2. Функціональні реалізації
- 1.3. Типологія агентів
- 1.4. Субстанційні реалізації
2. Чи може машина бути моральним агентом?
3. Чотири «пробіли»: проблема відповідальності

Лекція 3. Ризики та виклики штучного інтелекту. Відповідаючи на виклики, знижуючи ризики.

1. Класифікуючи ризики штучного інтелекту
2. Загальна типологія.
- 2.1. Дискримінація та токсичний контент
- 2.2. Приватність 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.
3. Нормативне регулювання штучного інтелекту в світі
- 3.1. ЄС
- 3.2. США
- 3.3. Великобританія
- 3.4. Китай
- 3.5. ООН, G7 та інші наддержавні актори

- 3.6. Регулювання ІІІ в Україні
- 4. Державні стратегії AI-Governance
 - 4.1. Німеччина
 - 4.2. США
 - 4.3. Китай
 - 4.4. Африка
 - 4.5. Південно-Східна Азія
- 5. Перспективи ІІІ у сфері GovTech
- 6. Методи протидії етичним ризикам та викликам
 - 6.1. Інституційні методи
 - 6.2. Технічні методи
 - 6.3. Гібридні підходи
 - 6.4. Чому проблеми не вирішено: trade-offs імплементацій, неформальний зміст вимог та бажань
- 7. ІІІ, економіка, робота майбутнього

Лекція 4. Роблячи ІІІ 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 5. Принципи безпечного дизайну систем ІІІ

- 1. Безпечний дизайн як метаетичний фреймворк
- 2. Принципи безпечного дизайну
 - 2.1. Резервування
 - 2.2. Прозорість
 - 2.3. Розподіл обов'язків
 - 2.4. Принцип мінімуму привілеїв
 - 2.5. Відмовостійкість
 - 2.6. Антикрихкість
 - 2.7. Механізми негативного зворотного зв'язку
 - 2.8. Глибинний захист
 - 2.9. Одомашнювання
 - 2.10. Приписуваність
 - 2.11. Переривність
 - 2.12. Агентична розбірливість
 - 2.13. Методи контролю здібностей
 - 2.14. Контейнеризація
- 3. Принципи в синергії: три приклади застосування

Лекція 6. Публічний ІІІ як альтернатива пропріетарності

- 1. Екосистема публічного ІІІ
- 2. ІІІ в муніципальному управлінні
- 3. Публічна підзвітність та нагляд
- 4. Стейкхолдери

Лекція 8. Підходи до створення відповідального та пояснюваного ШІ

1. xAI та підходи до його імплементації
2. Responsible AI та шляхи його імплементації
3. Легалістський підхід до реалізації безпечного ШІ

Лекція 8. Формування цінностей: проблема та підходи до вирішення

1. Проблема імплементації цінностей.
2. Методи:
 - 2.1. Еволюційні алгоритми.
 - 2.2. Асоціативне накопичення цінностей.
 - 2.3. Шаблони мотивації (motivation templates).
 - 2.4. Цінності як предмет дослідження та вивчення.
 - 2.5. Модуляція емуляції.
 - 2.6. Безпосереднє представлення.
 - 2.7. Інституційне впровадження та структурування.

Семінарські заняття

Основним завданням циклу практичних 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. Практичні заняття мають сприяти кращому засвоєнню теоретичного матеріалу з курсу «Етика Штучного Інтелекту». З метою інтеграції теоретичної компоненти (знання) із практичною (застосування знань) питаннями до практичного заняття є кейси за тематикою лекцій, за якими студент готує та презентує під час заняття доповідь, у якій представляє як своє бачення етичної проблеми зі сфери ШІ, так і варіант вирішення, який він вважає найкращим із можливих.

Семінар 1.

Тема: Штучний інтелект як проблема етики.

Завдання 1. Звернемося до актуальної на сьогоднішній день проблеми вузьких ШІ – безпілотного автотранспорту. Припустимо наявність нової моделі, розробленої таким чином, що її налаштування дозволяють власникові *обрати*, яким чином діятиме автомобіль у випадку аварійної ситуації ексклюзивно-диз'юнктивного типу (на відміну від інклюзивної диз'юнкції, в даному випадку: або А, або В, але ніяким чином не А та В одночасно), обираючи, в чіх інтересах діяти: намагатись врятувати свого пасажера чи пішоходів / інших пасажирів чи водіїв тощо – «іншу сторону». Як, в даному контексті, розподіляється відповідальність у випадку виникнення аварійних ситуацій? А саме: чи є розробник машини повністю відповідальним – за те, що створив саму можливість обирати, і в одному з випадків – не на користь випадкових людей, яким трапилася аварія з безпілотним авто? Чи, можливо, в момент вибору поведінки машини в налаштуваннях, слід перенести всю відповідальність на власника-пасажера, котрий, обираючи стратегію «рятувати мене» фактично інструктує машину «не рахуватися з життями інших» на свою користь? Але ж чи не є зворотнє питання не менш неприємною опцією, що виражає зневагу до життя індивіда: «За будь-яких обставин, рятуй інших та нехтуй моїм життям»? Можливо, відповідальність розподілена, до певної міри, між розробником та власником авто? Чи є варіант взагалі не створювати подібних опцій – головною мірою *саме для того*, щоб зняти моральну чи навіть юридичну (кримінальну) відповідальність? Обґрунтуйте ваші міркування з даних питань.

Завдання 2. Уявіть майбутнє, в якому ШІ були створені як AGI, інтегровані у суспільство та виконують в ньому критичні ролі на рівні з людьми – в уряді, освіті, охороні здоров'я тощо. Виникає громадський рух, учасники якого вимагають визнати цих ШІ як повноправних громадян-людей, включаючи права займати громадські посади офіційно, голосувати тощо (аж до альтернативних, в залежності від типу імплементації, варіантів служби та кар'єри у військовій

сфері). Вам поручено очолити комісію, що має дослідити та оцінити імплікації та аргументи цього громадського руху та надати експертну оцінку із рекомендаціями вдовольнити вимоги руху – повністю чи частково – або відмовити спільноті в їхніх вимогах. Чи підтримаєте Ви такого роду повноцінне визнання? Чи може таке рішення викликати занепокоєність щодо впливу ІІІ-громадян на людське суспільство та потенціальні зсуви у сфері владних структур. Наскільки такі занепокоєння, і які саме, Ви вважаєте обґрунтованими? Якщо слід відмовити ІІІ в такому праві, аргументуйте, чому саме.

Завдання 3. Уявіть, що Ви очолюєте групу консультантів з етики для уряду, який планує впровадження повністю автоматизованої (на основі ІІІ) системи правосуддя: від слідства до винесення вироків. Система обіцяє високу швидкість та майже абсолютну відсутність помилок. Але під час тестів виявляється, що в деяких складних або морально суперечливих випадках алгоритм пропонує розв'язки, котрі люди сприймають як «нелюдяні» (байдужі до унікальних обставин, співчуття тощо). Яким чином, на Вашу думку, слід врівноважити потребу в ефективності й швидкості правосуддя з необхідністю зберегти етичний вимір людського підходу до винесення вироків? Чи достатньо додаткового «модуля емпатії», або гібридної моделі «ІІІ + людина»? Які чинники можуть ускладнити такий підхід і як Ви пропонуєте подолати ці ускладнення?

Семінар 2.

Тема: Штучні моральні агенти та пробіл у відповідальності. Ризики та виклики штучного інтелекту.

Завдання 1. Припустимо, що проблеми доступу, змісту та контексту, притаманні сьогодні ІІІ-асистентам, можна буде вирішити вже найближчим часом. Але, як і будь-яке наше рішення та дія, вирішення проблем буде не без наслідків. Виходячи з сутності трьох проблем та їх експлікації, поміркуйте, до яких інших потенційних проблем у роботі ІІІ-асистента та його взаємодії з користувачем можуть в принципі призвести (і внаслідок чого) вирішені проблеми доступу, змісту та контексту? Обґрунтуйте Вашу думку. Як висновок, оцініть ваші міркування заключенням, чи є ці проблеми або ускладнення ціною, яку варто заплатити, оскільки кінцевий результат за функцією корисності переважає негативні ефекти.

Завдання 2. Уявімо собі три потенційних імплементації етичного «модулю» ІІІ, а саме: ІІІ-утилітариста; ІІІ-консеквенціаліста; ІІІ-деонтолога. Послугуючись 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 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 глузду" (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 глузд 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 4. Уявімо наступне. Ви є членом передової команди з дослідження та розробки ІІІ, 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 ІІІ до рекурсивного самовдосконалення? До яких наслідків може призвести відмова чи згода імплементування? Чи може в даному випадку, враховуючи феномен та проблему «зловісного ІІІ» як результату РС, існувати третя позиція, відмінна від однозначного схвалення чи відмови?

Завдання 5. Ви – лідер команди, яка займається розробкою дуже просунутої системи ІІІ, здатної генерувати гіперреалістичні діпфейк-відео. Відео такого роду можна використовувати на благо людей, створюючи фільми чи презентації, візуалізуючи ідеї тощо (все, що наразі роблять генеративні ІІІ із відео, тільки такі, що їх якість не відрізнити від реальної зйомки). Але звісно, технологія такого типу та можливостей має потенціал для використання і в недобрих чи відверто злочинних цілях та діях, від введення в оману до харасменту та шантажу. Усвідомлюючи це, як лідер проекту, чи продовжите Ви розвиток такої технології? Якщо так, яким чином ви зможете впевнитися у виключно етичному використанні технології, які запобіжні міри ви імплементуєте, які підготуєте запобіжні заходи?

Завдання 6. Уявімо ІІІ, 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 ІІІ реально сприяє науці більше, ніж середньостатистичний людський дослідник, чи не дискримінація це? Як вирішити конфлікт між традиційним уявленням про «науковця» та реальністю, де ІІІ може перевершувати людей у науковій творчості?

Семінар 3.

Тема: Відповідаючи на виклики, знижуючи ризики.

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 ІІІ, очолює Дослідницький Інститут Машинного Інтелекту) так переймається проблемою Василіска?

Завдання 2. Розгляньте гіпотетичне майбутнє, в якому люди створили AGI і він не вийшов з-під контролю і не став екзистенційною загрозою для людства. Тоді, мабуть, ІІІ стане фактором вирішення багатьох глобальних проблем, із якими ми стикаємося зараз. Згадайте інші, окрім суперінтелекту, екзистенційні ризики та загрози, і аргументовано перелічіть ті з них, із якими ІІІ безумовно і перш за все допоможе нам впоратися (яким чином він нам допоможе – відповідати необов'язково, лише за бажанням, на Ваш розсуд).

Завдання 3. Існує підхід до розробки ІІІ відомий як Comprehensive AI Services (Всеохопні сервіси ІІІ). Беручи за вихідний принцип розуміння ІІІ виключно як спеціалізованих інструментів, прибічники підходу закликають до поглиблення в дослідженнях та подальшу розробку вузьких ІІІ, 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 ІІІ 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 ІІІ. Самі вузькі ІІІ 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 ІІІ. 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 прогало б?

Семінар 4.

Тема: Роблячи ІІІ Етичним.

Парадигма публічного ІІІ як альтернатива пропріетарності.

Завдання 1. Чи вирішить проблему контролю рішення Стюарта Рассела, а саме – відсутність у AGI постановки мети та неможливість її мати? Які етичні колізії та проблеми постають разом із таким рішенням – для людства та для ІІІ? Що, як в ході власного розвитку, ІІІ зможе обійти наші обмеження на наявність мети? Яким може стати відношення ІІІ до людей, якщо ІІІ збагне, що саме його розробники є причиною початкової відсутності в нього мети? Яким чином можна запобігти в такому випадку війні людства та машин?

(Готуючи це питання, замислитесь, чи мають взагалі імplementовану «за замовчуванням» мету люди, коли народжуються?)

Завдання 2. Ви – суддя у справі, де корпорація створила ІІІ здатний до самозахисту в суді, аргументуючи на користь його права існувати та діяти незалежно від людей чи інших спостерігачів, регуляторів тощо. Корпорація, натомість, розглядає цю ІІІ-персону, лише як корпоративну власність, на підставі чого вимагає надати їй право знищити ІІІ, якщо він перестане слугувати цілям корпорації. Яке саме рішення і на чію користь приймете Ви у цій справі? Які фактори вплинуть на рішення і якого роду співвідношення прав Ви встановите у справі ІІІ проти корпорації-розробника?

Завдання 3. Ось яка справа. Корпорація створила ІІІ з достатнім рівнем самосвідомості, що він здатний до самозахисту в суді. ІІІ 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 ІІІ-персону, лише як корпоративну власність, на підставі чого вимагає надати їй право знищити ІІІ, якщо він перестане слугувати цілям корпорації.

В цьому світі, на відміну від світу з віртуальними працівниками, де дозволити собі з ними взаємодіяти могли лише найбагатші та юридичні а не фізичні особи (через що широка суспільна думка не була ані за, ані проти вбивства віртуальних працівників), справа набула розголосу, люди всіляко стають на бік ІІІ, бойкотують корпорацію, влаштовують маніфестації під центральним офісом, вимагаючи свободи штучної особистості. Суддя на слуханнях, хоча й завуальовано, але також експліцитно висловлює підтримку ІІІ. Єдині, кому не поталанило в цьому світі — це Ви, бо тут ви.... представляєте інтереси Корпорації, є їхнім юристом в суді по цій справі, адвокатом диявола, в якого щоразу летить капуста та невеликі камінці на шляху до зали суду через “коридор ганьби” під будівлею суда... Тож спробуйте представити більш чи менш (наскільки можливо) когерентні аргументи на користь того, що це все ще не особистість, а корпоративна власність, що Корпорація має владу над її життям та смертю. Знищення ІІІ 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. Спробуйте..

Завдання 4. Уявіть, що існує ІІІ, який здатен давати абсолютно вірну оцінку та судження щодо індивіда-особистості, із математичною точністю передбачаючи певні схильності та потенційні дії індивіда (роблячи це як мінімум так само, або краще, ніж людина-експерт з тих чи інших оцінок

Ваших дій). Знаючи, що в будь-якому разі, при проходженні експертизи, Ви отримаєте одну й ту саму оцінку, кому б Ви довірили її надати – людині чи машині? Чи, можливо, для Вас немає принципової різниці, якщо результат один і той самий? Обґрунтуйте Вашу позицію.

Завдання 5. Уявіть, що ви є частиною міжнародного комітету-робочої групи, перед якою стоїть завдання розробки етичних регулятивів та інструкцій стосовно використання ІІІ у бойових діях (та загалом в умовах війни). Було створено нову систему автономної зброї, котра здатна ідентифікувати конкретні цілі (рід військ (із розрізненням своїх / чужих навіть за 99% збігу, по конкретній тактичній деталі, що може слугувати критерієм такого розрізнення, зводячи нанівець ймовірність дружнього вогню, техніка, конкретний екземпляр ворожої техніки, який ідентифікується по тактичних чи інших знаках чи індивідуальних написах на техніці, якщо вони відомі власникам автономної зброї та їх було завантажено у dataset) та знищувати їх без усякого втручання в процес людини. Однак, на момент роботи вашої групи, існує небезпідставне занепокоєння щодо можливостей *абсолютно точного розрізнення* між цивільними та комбатантами при скупченні обох категорій в одному місці у великій кількості. Як член комітету, чи дозволите ви загалом використання таких систем, враховуючи, що на даний момент це лише результати розрахунків та передбачення, а реального випадку вбивства автономною системою цивільних, на щастя, не було. Які умови та обмеження можна накласти на такого роду зброю для забезпечення етичного її використання?

Завдання 6. Повернемося до двічі згаданої в лекції, важливої теми: обмеження наших бажань, які має задовольняти штучний інтелект. Де є, і чи є вона, лінія розмежування між “ще допустимим” для виконання бажанням людини і “вже неприпустимим”? За яких умов постає проблема нерівного доступу до ІІІ, чи можна і як змінити його на більш егалітарне користування. А чи слід це робити? Чи достатньо обмежитися оформленням у правовому полі відповідальності ІІІ на рівні з людьми? Чи не змусить це ІІІ, який заради користувача порушує закон просто робити те ж саме але більш вміло уникаючи наслідків? Чи маємо ми прийти до консенсусу в формі свідомого, розумного та всезагального самообмеження (“Жоден з людей та не забажає Х!”) щодо того, що може зробити для нас ІІІ? Чи це має бути підкріплено встановленням “планки” його можливостей (які можна і зірвати...).

Семінар 5.

Тема: Підходи до створення відповідального та пояснюваного ІІІ.

Завдання 1. Уявімо, що існує незліченна кількість цифрових (віртуальних) особистостей, формою реалізації та типом близькі до тих, можливість яких було розглянуто на лекції. Ви – власник компанії, у якій частина працівників – віртуальні. Кожен із них є повноцінною особистістю, зі своїм характером, звичками, вадами та чеснотами. Але в певний момент прибутки компанії падають через вихід нової моделі ІІІ, на архітектурі якого реалізовано більш ефективну, більш вдалу віртуальну особу (з точки зору її як працівника). Ви можете дозволити собі замінити нинішніх працівників на нових, але «звільнити» віртуальну особистість означає повністю стерти її та видалити – фактично, вбити. Чи змогли б ви таким чином «звільнити» (особисто вбити), скажімо, 34 віртуальних осіб таким чином задля того, щоб найняти 25 нових, більш ефективних працівників? Якщо ні, як би Ви вийшли з ситуації, котра склалася б для Вашого бізнесу в контексті конкуренції та ефективності? Які взагалі варіанти та рішення існують в даному контексті?

Завдання 2. Уявімо майбутнє, в якому ІІІ-сутності отримали певні права як цифрові громадяни, включаючи не просто право на безперешкодне існування, а й на розвиток, встановлення, рух до – власних цілей. Однак, запропоновано новий закон із метою контролю та регулювання, такий, що ІІІ-персону буде позбавлено всіх громадянських прав, якщо буде достеменно відомо, що цілі ІІІ-персони конфліктують з інтересами людей; позбавлені прав ІІІ отримують статус «просто інструментів», якими люди будуть послуговуватися виключно у власних цілях, не маючи мети і маючи сегментарні обмеження мислення. Як законотворець, чи підтримали б Ви такий закон? Як

слід примирити потенційний конфлікт між автономією *та* волею до неї в ШІ та людським бажанням контролю над ШІ; які етичні принципи матимуть вплив на Ваше рішення?

Завдання 3. Порівняйте агентність людини та ШІ з точки зору (ір)раціональності. Люди, ми знаємо, неповністю раціональні агенти, на відміну від, як вважається, ЗШІ. Які недоліки та переваги, в коротко-, середньо- та довгостроковій перспективі може нам дати наша ірраціональна сторона? В чому або коли бути абсолютно раціональним агентом краще? Як щодо ситуації, у якій в ШІ поступово, в ході машинної еволюції, виникають емоції і відчуття в тому числі ірраціональні? Що, на Вашу думку, відчуватиме ШІ — ультрараціональний агент — вперше зробивши якусь “дурню”, ірраціональну дію, щось, що він не може сам собі (тим паче іншим) обґрунтувати? Чи не стане тут в нагоді наш досвід в усій його суперечливості? Нас, що вчаться давати раду ірраціональному з малих літ?

Семінар 6.

Принципи безпечної розробки та підходи до формування цінностей.

Завдання 1. Уявіть себе в ролі очільника спеціальної комісії чи наглядової ради з етики ШІ, яка впроваджує правила використання та створення ШІ, а також окрім іншого виступає арбітром у конфліктних чи проблемних ситуаціях стосовно ШІ, які виникають. До Вас потрапила наступна справа: ШІ, розроблений спеціально для управління інфраструктурою міста-мегаполісу із великою індустріальною забудовою та великою кількістю промисловості різних типів, починає приймати управлінські рішення, котрі ставлять у пріоритет довгострокові плани та проекти збереження та стабілізації навколишнього середовища понад людською зручністю та комфортом «тут і зараз» (тим самим частково викликаючи дискомфорт у частини населення). Чи слід комісії підтримати автономію ШІ у його новій політиці прийняття рішень, або навпаки – прийняти та ввести в силу процедури насильного перепрограмування, котрі повернуть дії ШІ до узгодженості з людськими інтересами та перевагами. Як Ви обґрунтуєте своє рішення? Чи слід зробити його прецедентом, тобто першим випадком у своєму роді, що стане підставою для того, щоб у подальшому, при виникненні саме цієї конкретної проблеми, рішення приймалося за зразком першого випадку (у формі закону або конвенції)? Чи, все ж, ситуація не завжди може бути так само однозначною*, у в кожному приватному конкретному одиничному випадку, рішення може бути як за дії ШІ, так і проти?

Завдання 2. Візьміть одну з графічних моделей ризиків та систем безпеки із попередньої лекції (друга частина) та зробіть таку модель для одного з підходів до імплементації цінностей, або одного з тих, які розглянуто на лекції 7, або на Ваш особистий розсуд, якщо вважаєте, що є цікавіші підходи, варті уваги. Необов'язково 1 в 1 відтворювати мої візуалізації, адаптуйте так, як бачите за потрібне та за можливе, оскільки тут все ж маємо справу не з ризиками, а трохи інший предмет.

Завдання 3. Розглянемо майбутнє, в якому ШІ успішно працюють у якості HR'ів, використовуючись для оцінки кандидатів на вакантні посади на основі аналізу резюме, співбесіди та персональних даних пошукача. Загалом, ШІ продемонстрували більшу точність у передбаченні успішних наймів працівників (у порівнянні з людьми-рекрутерами). Однак, подальші дослідження свідчать про те, що при цьому, не дивлячись на ефективність, ШІ має схильність до вподобання певної демографічної групи як пріоритетнішої за всі інші – схильність до дискримінації. Дізнавшись про цю інформацію (і отримавши підтвердження її істинності), маючи змогу вплинути на глобальну реакцію на ці нові дані, чи рекомендували б Ви продовжувати використовувати ШІ в якості HR? Якщо так (з огляду на те, що рекомендації загалом є кращими, ніж людські), яким чином Ви вбачаєте вирішення проблем та питань егалітарного підходу до здобувачів, чесності, прозорості та принципу різності, наймаючи людину на роботу? Яким чином слід відрегулювати алгоритми передбачення та вибору в ШІ, аби уникнути дискримінації?

Семінар 7.

Тема: Проблема контролю: дивлячись в теперішнє і майбутнє.

Завдання 1. Яка з чотирьох можливих версій реалізації ІІІ здається для Вас найкращою чи найбільш оптимальною: ІІІ-оракул, джин, інструмент чи суверен (NB! Не «найбезпечнішою», а просто – найліпшим варіантом у Вашому розумінні)? Обґрунтуйте Ваш вибір та його критерії.

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

Завдання 3. Ви — головний інженер з питань та засобів безпеки у перспективному проєкті з розробки загального ІІІ. Спираючись на представлені в лекції методи, деталізовано представте свій план методів, дій, рішень, сукупність яких ви використаєте для максимізації поточного і подальшого контролю та убезпечення від небажаних сценаріїв, пов'язаних із ІІІ 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.

Для виконання експрес-контролю необхідно відвідувати лекції та семінари, самостійно опрацьовувати проблемні питання.

Самостійна робота здобувача

Засвоєння змісту дисципліни «Етика Штучного Інтелекту» разом із аудиторними заняттями передбачає проведення здобувачами індивідуальної роботи з метою самоконтролю знань та підготовки до занять. Систематична індивідуальна робота дає можливість закріпити матеріал курсу, акцентує увагу на головних проблемах тем, що вивчаються. В межах курсу передбачені такі види самостійної роботи: підготовка до аудиторних занять за змістом прослуханої лекції із використанням додаткових джерел, головним чином – базової та допоміжної літератури (виконується протягом тижня після лекції, перед кожним практичним заняттям); виконання домашнього завдання (теми домашніх завдань пропонуються на другій лекції; домашнє завдання має бути виконано і здано на перевірку не пізніше, ніж за два тижні до дня проведення заліку); підготовку до модульної контрольної роботи (проводиться протягом однієї академічної години на аудиторному занятті під час першого календарного контролю (атестації)); підготовку до залікової співбесіди, котра є однією з двох частин заліку з дисципліни (має місце протягом семестру в процесі отримання, здобуття та опанування теоретичного знання, методології дослідження та практичних навичок в межах курсу).

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 на одне комплексне питання у вигляді міні-есе. Приклад завдань модульної контрольної роботи наведено у додатках до даного силабусу.

№ з/п	Самостійна робота студентів	Кількість годин
1	Підготовка до аудиторних занять	20
2	Підготовка до складання модульної контрольної роботи	4
3	Підготовка до заліку	6
	Всього	30

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Відвідування занять

Правила відвідування занять. Відвідування занять є вільним, бали за присутність на лекції та семінарських заняттях не додаються, штрафні бали не передбачаються. Однак, студентам рекомендується відвідувати заняття, оскільки на них викладається теоретичний матеріал та розвиваються навички, необхідні для отримання певних позитивних результатів вивчення дисципліни завдяки активній участі у роботі на семінарських заняттях.

Дистанційний режим навчання. У разі запровадження обмежень на відвідування університету у разі організації освітнього процесу у змішаному /дистанційному режимі пов'язаних з введенням режиму воєнного стану в державі (або карантину), освітній процес 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>). У режимі дистанційного навчання заняття відбуваються у вигляді онлайн-конференції на платформі Zoom. Посилання на конференцію видається на початку семестру. З метою забезпечення якісної підготовки здобувачів, дистанційний курс дисципліни розміщено на Платформі дистанційного навчання «Сікорський» (<https://www.sikorsky-distance.org>). Результати оцінювання висвітлюють у АС «Електронний кампус» на особистій сторінці здобувача (<https://ecampus.kpi.ua>).

Правила поведінки на заняттях. На заняттях слід дотримуватись норм етичної поведінки визначених у Кодексі честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (<https://kpi.ua/code>). В аудиторіях на лекційних та семінарських заняттях, а також під час проведення занять у формі відеоконференції, користуватись мобільним телефоном потрібно у беззвучному режимі. Під час дистанційного режиму навчання слід використовувати інформацію розміщену на Платформі дистанційного навчання «Сікорський» / Електронний кампус.

Позааудиторні заняття та залучення професіоналів-практиків. Під час вивчення дисципліни можливі позааудиторні заняття, що включають відвідування науково-практичних заходів, лекцій, тренінгів в межах тематики дисципліни.

Визнання результатів навчання, набутих у неформальній/інформальній освіті. Зарахування сертифікатів проходження дистанційних чи онлайн курсів здійснюється згідно Положення про визнання в КПІ ім. Ігоря Сікорського результатів навчання, набутих у неформальній/інформальній освіті (затверджено та уведено в дію наказом від 09.05.2023 р. № НОН/157/2023).

Правила призначення заохочувальних та штрафних балів. Заохочувальні бали можуть бути отримані за участь в олімпіадах (всеукраїнських, університетських), конкурсах наукових робіт, наукових (науково-практичних) конференціях (10% від суми рейтингових балів). Заохочувальні бали не входять до основної шкали РСО. Штрафні бали не передбачаються.

Політика оцінювання контрольних заходів Оцінювання контрольних заходів здійснюється згідно з Положенням 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/index.php/node/32>). Нижня

межа позитивного оцінювання кожного контрольного заходу складає не менше 60% від балів, визначених для цього контрольного заходу, а негативний результат оцінюється в 0 балів.

Поточний контроль проводиться впродовж семестру з метою забезпечення зворотного зв'язку між викладачем і здобувачами у процесі навчання та для перевірки рівня теоретичної та практичної підготовки здобувачів на кожному етапі вивчення навчальної дисципліни.

Календарний контроль проводиться двічі на семестр і передбачає проведення модульної контрольної роботи (поділеної на 2 частини), яка здійснюється у вигляді письмової контрольної роботи. Умовою отримання позитивної оцінки з календарного контролю з навчальної дисципліни є значення поточного рейтингу здобувача не менше, ніж 50 % від максимально можливого на час проведення такого контролю.

Семестровий контроль проводиться у формі заліку за РСО-1, що передбачає оцінювання результатів навчальної діяльності здобувача впродовж семестру, передбачених заходами поточного контролю. Результати оцінювання висвітлюються у АС «Електронний кампус» на особистій сторінці здобувача (<https://ecampus.kpi.ua>).

Політика дедлайнів та перескладань. МКР складаються лише у призначений день. Якщо контрольні заходи, або виконання завдань пропущені з поважних причин (хвороба або вагомі життєві обставини), надається можливість додатково скласти завдання протягом двох найближчих тижнів. Невиконання завдань, а також порушення термінів їх виконання з неповажних причин не надасть можливості набрати відповідні бали рейтингу. 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, 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>).

Процедура оскарження результатів контрольних заходів. У випадку незгоди здобувача 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 щодо академічної доброчесності. Необхідним під час виконання завдань з дисципліни є дотримання політики та принципів академічної доброчесності (<https://kpi.ua/academic-integrity>), які, у тому числі викладено у Кодексі честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (<https://kpi.ua/code>), **Положенні про систему запобігання академічному плагіату в КПІ ім. Ігоря Сікорського** (<https://osvita.kpi.ua/node/47>). У разі виявлення дублювання виконаних завдань, плагіату роботи отримують нульовий рейтинг.

Політика використання штучного інтелекту. Використання штучного інтелекту регламентується «Політикою використання штучного інтелекту для академічної діяльності в КПІ ім. Ігоря Сікорського» (<https://osvita.kpi.ua/node/1225>).

Здобувачі вищої освіти можуть використовувати інструменти моделей генеративного ШІ для: проведення дискусій, обговорення ідей, розвиваючи власні думки та аналізуючи ідеї «співрозмовника»; перегляду своїх ідей, їх переформулювання для поліпшення структури та їх чіткого, більш конкретного висловлювання; організації самостійного навчання, розвитку навичок письма, вдосконалення стилю висловлювання та розширення лексичного запасу; автоматизованої перевірки граматики, структури та стилю своїх написаних робіт; створення мультимедійних матеріалів (анімації, відео чи інтерактивних презентацій); розробки ігрових сценаріїв, що сприяє розумінню різних ситуацій та розвитку стратегічного мислення; аналізу великих обсягів даних у своїх дослідженнях.

Здобувачам не рекомендується покладатися на ШІ як на єдине джерело інформації. Важливо перевіряти та аналізувати отримані дані з інших авторитетних джерел. Усі випадки використання ШІ для виконання завдань мають бути чітко вказані та задокументовані. Це стосується як використання текстових генераторів, так і інших інструментів ШІ. Використання ШІ має відповідати принципам академічної доброчесності. Недотримання цього положення розглядатиметься як порушення академічної етики.

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

Семестровий контроль з дисципліни «Етика штучного інтелекту» передбачений у вигляді заліку, 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 і штрафних 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	60	10	6	60
2.	Підготовка есе	12	12	1	12
3.	Експрес-контроль	12	12	1	12
4.	МКР (2 год.)	16	16	1	16
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 – 10. Максимальна кількість балів на всіх семінарських заняттях дорівнює $10 \text{ балів} \times 6 \text{ видів робіт} = \mathbf{60 \text{ балів}}$.

До видів робіт відносяться:

- Відповіді на семінарських заняттях (презентація – індивідуальна або в парі – одного на вибір семінарського питання у вигляді кейсу, який презентують у постановці проблем та пропонують варіант аргументованого рішення). Ваговий бал – 10. Кількість відповідей – **4**. (Максимальна кількість балів - 40).
- Участь у обговоренні питань семінарів. Ваговий бал – 10. Кількість відповідей – **2**. (Максимальна кількість балів - 20).

Чотири рівні оцінювання:

10 балів – відповідь повна, чітка, логічно структурована, містить глибоке розуміння теми, підтверджується аргументами, прикладами або цитатами. Студент відповідає на уточнюючі питання викладача. Відсутні помилки.

9-8 балів – відповідь майже повна, добре структурована, але можуть бути незначні неточності або недостатньо глибокий аналіз.

7-6 балів – відповідь поверхнева, може містити помилки або недоліки в аргументації чи структурі.

0 балів — відповідь відсутня або не відповідає питанню (менше 60% потрібної інформації).

2. Підготовка есе:

На початку курсу слухачам пропонується опрацювання одного з додаткових джерел на вибір і виконання домашнього завдання у формі есе, яке є рефлексією на проблематику джерела, або безпосередньою реакцією на зміст прочитаного. Ваговий бал - **12**. Загальна кількість Домашніх завдань протягом курсу: 1 (крайній термін здачі робіт – до початку другого Календарного контролю). Максимальна кількість балів – 12.

12 балів – есе повністю відповідає заявленим темі та меті, є внутрішньо несуперечливим, логічно консистентним та когерентним, аргументи валідні та достатнім чином підкріплені і є переконливими; додатково можуть бути наведені контраргументи щодо позиції автора (та спростовані відповідним чином).

11-10 балів – есе є добре обґрунтованим, логічно когерентним та консистентним, аргументи є послідовними та відображають предмет аргументації, проте частина не є релевантною, а деякі змістові компоненти виграли б від уточнень; або есе написано добре, проте зміст частково розходиться з декларативними постулатами на початку роботи; або аргументи інтуїтивно є валідними, проте недостатнім чином експліцитно обґрунтованими.

9-8 балів – есе містить необґрунтовані імплікації, містить значні «пробіли» в аргументації, містить логічні помилки або вдається до логічних еквівокацій, є змістовно незавершеним або відсутня кореляція між змістовними компонентами есе (висновків із основною частиною, заявленої мети або імплікацій – аргументам та висновкам тощо).

7 балів – “достатньо”, – есе носить декларативний та поверхневий характер і є фактично доксографією (зібранням цитат із чужими думками з повною відсутністю власних міркувань).

0 балів – “незадовільно”, тобто або дана робота відсутня, або повністю не відповідає поставленому завданню, але здана.

3. Проведення експрес-контролю:

В ході проведення першого Календарного контролю, слухачам пропонується виконання експрес-тесту, що складається з 12 тестових питань за матеріалом та темами Лекцій 1-4, Семінарів 1-4. Тип питань: 6 питань типу «Вибір однієї правильної відповіді» та 6 питань типу «True / False». 1 правильна відповідь на питання = 1 бал, відповідно до чого Ваговий бал та Максимальна кількість балів – **12**. Загальна кількість Експрес-контрольних протягом курсу: 1 (відповідно до дат першого Календарного контролю). Максимальна кількість балів – 12. Оцінювання експрес-контролю базується на підрахунку правильних відповідей на 12 питань вказаних типів.

4. Виконання модульної контрольної роботи:

Ваговий бал - **16**. Загальна кількість МКР протягом курсу: 16 (МКР проводиться під час другого Календарного контролю). Максимальна кількість балів – 16.

16 балів – 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.

15-14 балів – 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.

13-12 балів – не повна відповідь на всі поставлені питання, що свідчить про правильне розуміння суті питання, ознайомлення студента з матеріалом лекцій та підручника; неточності у відповідях.

11-10 балів – відсутність відповіді на певні питання, що свідчить про поверхове ознайомлення студента з навчальним матеріалом або значні похибки у відповідях.

0 балів – відповідь відсутня або не відповідає питанню (менше 60% потрібної інформації).

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 балів	≥ 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).

Штрафні бали передбачені за відмову від відповіді на контрольні запитання з теми семінару і невиконання запропонованих на семінарському занятті контрольних завдань (експрес-опитування, тестів). Заохочувальні та штрафні бали не входять до основної шкали РСО, а їхня сума не може перевищувати 10% рейтингової шкали.

Підсумковий контроль: ЗАЛІК

Підсумковий контроль проводиться відповідно до навчального плану у вигляді заліку в терміни, передбачені встановленим графіком навчального процесу. Форма проведення семестрового контролю комбінована і складається з двох частин. Першою є написання есе на одну з запропонованих тем; максимальна кількість балів, передбачена за виконання даної частини заліку складає 25 балів. Другою частиною є – залікова співбесіда за трьома питаннями залікового білету, котрий здобувач тягне на початку заліку.

Кожне питання оцінюється у 25 балів за такими критеріями:

- 25 балів – відповідь на питання викладено правильно, всебічно, безпомилково і логічно;
- 21-24 балів – достатньо повна відповідь, або незначні неточності;
- 17-20 балів – достатньо повна відповідь, або незначні неточності;
- 13-16 балів – неповна відповідь, деякі помилки;
- 0 балів – відповідь відсутня, або є цілком невірною, містить менше, ніж 60% необхідної інформації.

Здобувач отримує позитивну залікову оцінку за результатами роботи протягом семестру, якщо має підсумковий рейтинг за семестр не менше 60 балів та виконав умови допуску до семестрового контролю.

Умови допуску до заліку: рейтинг ≥ 36 б. 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.

Не виконані умови допуску → Не допущено.

< 60 балів → залікова к/р +співбесіда.

≥ 60 балів = оцінка (відмінно, дуже добре, добре, задовільно, достатньо, незадовільно). Оцінка може бути підвищена за бажанням за рахунок виконання залікової к/р +співбесіда.

Залік проводиться в період останніх двох тижнів теоретичного навчання у семестрі, як правило, на останньому за розкладом занятті з навчальної дисципліни «Етика Штучного Інтелекту». Результати контрольних заходів доступні до ознайомлення авторизованим користувачам в їх особистих кабінетах автоматизованої інформаційної системи «Електронний кампус».

Принцип визначення підсумкової оцінки. Рейтингова оцінка доводиться до здобувачів на передостанньому занятті з дисципліни в семестрі. Здобувачі, які виконали всі умови допуску до заліку і мають рейтингову оцінку 60 та більше балів, отримують відповідну до набраного рейтингу оцінку без додаткових випробувань.

Зі здобувачами, які виконали всі умови допуску до заліку та мають рейтингову оцінку менше 60 балів, а також з тими здобувачами, хто бажає підвищити свою рейтингову оцінку, на останньому за розкладом занятті з дисципліни в семестрі викладач проводить семестровий контроль у вигляді залікової контрольної роботи (письмова) + співбесіда.

Максимальна сума балів складає **100**.

Сума балів переводиться у систему оцінювання згідно з таблицею.

Таблиця переведення рейтингових балів до оцінок за університетською шкалою

<i>Кількість балів</i>	<i>Оцінка</i>
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.

В умовах дистанційного режиму організація освітнього процесу здійснюється з використанням технологій дистанційного навчання: система Електронний кампус. Для більш ефективної комунікації з метою розуміння структури навчальної дисципліни і засвоєння матеріалу використовуються сервіси для організації онлайн-конференцій та відеозв'язку (наприклад, «Zoom»), електронна пошта, месенджери.

Навчальний процес у дистанційному режимі здійснюється відповідно до затвердженого розкладу навчальних занять.

Робочу програму навчальної дисципліни (силабус):

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

Ухвалено на засіданні кафедри філософії (протокол № 11 від 27.01.2026 р.).

Погоджено Методичною радою університету (від 26.06.2026 р.)

ДОДАТОК А. Модульна контрольна робота (Зразок)

Частина І.

- 1). Якими були причини та передумови «другої зими ШІ»?
- 2). Визначте параметр / властивість агентичності (agenticness) та зазначте його відмінність від агентності (agency).
- 3). Перелічіть та коротко охарактеризуйте всі основні компоненти телеології ШІ.
- 4). Поясніть сутність репрезентації етики ШІ як *спекулятивної* етики.
- 5). У який спосіб пропонується імплементувати шаблони мотивації у AGI?

Частина ІІ.

Поміркуйте над проблемою критеріїв, за якими можна стверджувати, чи є той чи інший ШІ загальним. Чи достатніми є два домінуючих критерії, а саме: математика та логіка, покрокове розв'язання задач із яких, супроводжуючись поясненнями кожного кроку, є свідченнями наближення до ЗШІ або навіть його створення. Які ще критерії Ви вбачаєте доцільними в якості «індикаторів» переходу від вузького до загального ШІ? Чи існують наразі критерії, котрі допоможуть нам виявити «зловісний» (malevolent) ШІ? Якщо не існують, чи можливо їх створити взагалі, і наскільки ефективними вони потенційно можуть бути? Чи є проблема критеріїв підпроблемою інших, «великих» проблем (проблема контролю, проблема імплементации тощо), або є аж ніяк не меншою від них, серйозною проблемою самою по собі.

ДОДАТОК Б. Приклади та зразки питань для залікової співбесіди, тестових завдань та тем для написання есе

I. Питання для залікової співбесіди

- 1) Визначте та порівняйте два методи формування цінностей у ЗШІ, а саме: шаблони мотивації та модуляцію емуляції.
- 2) Дайте визначення «суперінтелекту» за Ніком Бостромом. Зазначте та схарактеризуйте три форми суперінтелекту.
- 3) Розкрийте сутність, зміст та значення явища та концепту «інтелектуальний агент» та його роль у дослідженнях та розробці ШІ.
- 4) Визначте та порівняйте два методи формування цінностей у ЗШІ, а саме: асоціативне накопичення цінностей та цінності як предмет дослідження та вивчення.
- 5) Перелічіть основні причини та передумови, за яких ШІ може стати екзистенційним ризиком.

II. Темы для написання есе

1) Один із варіант майбутнього – AGI, після його створення, реплікується до нескінченності, а саме – до кількості рівної населенню планети. Кожна людина відтепер має персонального асистента, імплементованого у бажаній для кожної людини формі (робот, додаток, чатбот, кіборг тощо). Величезну кількість проблем людства вирішено завдяки суперінтелекту, котрий використовує міць свого інтелекту аби допомогти людям досягти їх цілей. Але люди залишаються людьми, і далеко не кожна людина бажає оточуючим лише добра... Уявімо деякого Х, метою якого є завдання оточуючим витончених (по-садистськи) страждань. Вочевидь, Х використовуватиме свого асистента ШІ для цієї мети: або непрямим способом, спостерігаючи, як асистент завдаватиме іншим страждань, або складатиме для Х плани, як заподіяти шкоду оточуючим, обходячи законодавство та інші потенційні контрміри.

Як, завдяки і за рахунок чого маємо дати раду проблемі злих намірів? Яким чином утримати / обмежити AGI-помічників так, щоб вони не могли заподіяти лиха оточуючим? Чи достатніми будуть обмеження етичного змісту як частина початкової етики ШІ, чи не призведуть вони до обмеження функціоналу і, як результат, зменшення рівня інтелекту у ШІ? Чи можливо взагалі зробити в цьому випадку щось саме з ШІ, чи вирішення проблеми – виключно в людині? Обґрунтуйте ваші міркування з цього приводу.

2) Згадайте запропоноване Стюартом Расселом вирішення проблеми контролю: ШІ завжди має ставити в пріоритет людські цілі над будь-якими потенційно самовизначеними цілями, навіть якщо ШІ буде повноцінною особистістю, самосвідомою та виокремленою для себе від інших та світу як такого. Вашим завданням є презентувати контраргумент у дебатах із Расселом, захищаючи ідею того, що ШІ має мати право розвиватися і, відповідно, переслідувати також і власні цілі, навіть якщо вони можуть не завжди співпадати з людськими. Які аргументи Ви, на вашу думку, можете використати на противагу погляду Рассела, у який спосіб ви підтримали б ідею автономії ШІ?

3) Недалеке майбутнє, в якому системи ШІ інтегровані в юридичні процеси всіх типів, допомагаючи суддям визначати в тому числі й вирок, базуючись на ретельному аналізі даних минулих справ. Системи допомагають стандартизувати вирок, але, переходячи до таких узагальнень, існує небезпека того, що система нехтуватиме та проглядатиме нюанси індивідуальної справи, взятої окремо поза контекстом. Чи підтримуєте Ви використання ШІ в юридичній системі, повноцінне чи обмежене, в даному контексті? Як можна впевнитись у тому, що правосуддя залишається чесним та індивідуалізованим, при цьому отримуючи підтримку можливостями ШІ?

ДОДАТОК В. Тестові завдання для самостійного виконання з метою набору додаткових балів (Зразок)

1) Якої форми реалізації ІШ не існує?

- A) Генеративний ІШ
- B) ІШ-ідол
- C) ІШ-оракул
- D) ІШ-суверен

2) Метод контролю здібностей ІШ, при якому ІШ 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) Синглтон
- B) Сингулярність
- C) Однополярність
- D) Монототальність

4) Багатополярний сценарій –

- A) світовий порядок, протилежний попередньому визначенню.
- B) сценарій настання ери ІШ, в ході якого, внаслідок якого з'являються та співіснують кілька конкурентних загальних ІШ.
- C) підхід до проблеми визначення цінностей ІШ, при якому кожен цінність представляють через її протилежності, котрі в ній фіксуються.
- D) методологія створення ІШ, яка характеризується одночасним використанням декількох різних підходів до його реалізації.

5) У дослідженнях та розробці ІШ 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 ІШ має справу з _____ .

- A) сьогоdnішніми проблемами та викликами ІШ.
- B) етичними дилемами та колізіями майбутнього, неактуальними сьогодні.

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

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