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	***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	*Загальна кількість: 60 год.* *Лекційних занять: 18 год.* *Практичних занять: 18 год.* *Самостійна робота студентів: 24 год.*
Семестровий контроль/ контрольні заходи	*Залік, МКР*
Розклад занять	*http://rozklad.kpi.ua/Schedules/ViewSchedule.aspx?v=7ce805c3-c7b4-4cc7-954e-1245a862edeb*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: *Казаков Мстислав Андрійович* *Кандидат філософських наук* potya@ukr.net Практичні / Семінарські: *Казаков Мстислав Андрійович* *Кандидат філософських наук* kazakov.mstyslav@lll.kpi.ua

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчання 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 покладаємося 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, виграє літературний приз у Японії; подібно DeepBlue у шахах, AlphaGo виграє 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 ШІ та ШІ 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 ШІ, котра розвивається з невловимою швидкістю; метою курсу є дослідження етичних імплікацій в технологіях ШІ, 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 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 широкий спектр варіантів дій, вчинків, рішень, підходів, тактик та стратегій.

Програмні результати навчання.
Набуття здібностей та здатностей 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;
знання та вміння генерувати нові знання, ідеї, моделі та гіпотези щодо довгострокових та високоабстрактних моральних питань та 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.

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.
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 дослідження.

3. Зміст навчальної дисципліни

Тема 1. Штучний Інтелект, етика та їх інтерсекція.
Тема 2. Види ШІ: наявні та майбутні.
Тема 3. Алгоритмічні похибки, помилки та «галюцинації» ШІ.

Тема 4. ШІ в контексті екзистенційних ризиків.
Тема 5. Проблеми мети, сумісності та контролю.
Тема 6. Формування цінностей у Загального ШІ.
Тема 7. ШІ як спекулятивна етика: огляд потенційних рішень для AGI.
Тема 8. Етичний контекст форм реалізації та імплементації ШІ.
Тема 9. Агентичність та форми агентності ІШ майбутнього.

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. Коцовський В.М. (2016) Методи та Системи Штучного інтелекту. Конспект лекцій. – https://shorturl.at/E16cf
4. 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
5. Рассел С. (2020) Сумісний з людиною: Штучний інтелект і проблема контролю. Київ: Форс Україна
6. Шаров С. (2023) Сучасний стан розвитку штучного інтелекту та напрямки його використання. - https://www.researchgate.net/publication/370595289_Sucasnij_stan_rozvitku_stucnogo_intelektu_ta_naprramki_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.

Навчальний контент

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 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. 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

1.1. Штучний Інтелект: Історія ідеї, визначення, поточний стан справ (реалії та спекуляції).
1.2. Етика: прикладна філософська дисципліна про вчинки, взаємодію та мораль.
1.3. Етика ШІ як спеціалізована етика, її характеристичні особливості.
1.3.1. Виклики ШІ як виклик етиці.
1.3.2. Етика ШІ як єдина етика зі спекулятивною складовою.
1.3.3. Метаетичні імплікації в питаннях реалізації Загального Штучного Інтелекту.

Лекція 2. Загальний (Artificial General Intelligence) та «Вузький» Штучний Інтелект
1. Narrow AI та Artificial General Intelligence.
2. Інтелектуальні агенти та середовища.
3. Основні (теоретичні) способи реалізації AGI.
3.1. Емуляція мозку людини.
3.2. Стандартна модель штучного інтелекту.
3.3. Аугментація біологічного мозку.
3.4. Нейроінтерфейси «людина – машина».
3.5. Розумна велика мовна модель (LLM).
4. Актуальні проблеми «Вузького ШІ»:
4.1. Алгоритмічні похибки, упередження та індуковані галюцинації.
4.2. Недоліки доступу, змісту та контексту: три проблеми ШІ-асистентів.
4.3. Проблеми приватності та збору даних.
4.4. Безпілотний транспорт та Автономна зброя.

Лекція 3. Виклики загального ШІ

1) Етика ШІ як спекулятивна етика.
2) Кінетика інтелектуального вибуху.
3) Знищення людства заради досягнення мети.
4) Синглетон(и) та Глобальна стратегічна перевага.
6) Соціальні наслідки імплементації AGI: від нового масового безробіття до інфантилізації та техно-варваризму.

Лекція 4. Штучний Інтелект як Екзистенційний Ризик. Поняття Суперінтелекту

1) Катастрофічні та екзистенційні ризики
2) Загрози екзистенційного рівня:
2.1. Нанотехнології.
2.2. Біологічна та ядерна зброя.
2.3. Контингентні катастрофи.
2.4. Суперінтелект.
3) Концепт суперінтелекту.
4) Форми суперінтелекту:
4.1. Швидкий (Слабкий) Суперінтелект.
4.2. Колективний Суперінтелект.
4.3. Якісний Суперінтелект.

Лекція 5. Проблеми Мети та Контролю

1) Мета як характеристична риса інтелекту.
2) Структурні та функціональні компоненти телеології ШІ.
3) Теза ортоґональності та теза інструментальної конвергентності.
4) Неузгодженість мети із людськими цілями та проблема хибної інтерпретації.
5) Вирішення проблеми контролю за Стюартом Расселом.

Лекція 6. 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.

Лекція 7. Вирішення проблем майбутнього

1) Критерії відбору цінностей:
1.1. Непряма нормативність: ревізія:
1.2. Принцип епістемічної переваги.
2) Контроль здібностей:
2.1. Пастки.
2.2. Контейнеризація та її методи.
2.3. Пригнічення.
2.4. Методи заохочення.
3) Відбір мотивації (шляхи селективної імплементації етики):
3.1. Прямі (прескриптивні) етичні вимоги.
3.2. Непряма нормативність.
3.3. «Одомашнення».
3.4. За «принципом доповнюваності» (до «зерна» етики інтелектуального агента).
4) Когерентне екстрапольоване бажання.

Лекція 8. Етичні виклики реалізації AGI.

1) Антропоцентричні похибки та кореляціоністські імплікації.
2) Правовий та соціальний статус майбутніх штучних чи віртуальних індивідних інтелектів.
3) Проблема фундації: між «Зерном ШІ» та «Карнаповим роботом».
4) «Дегуманізовані» ШІ як парадигма етосу «людина – ШІ»: від «машин, створених служити нам» (Рассел) до «машинної приватної власності»
5) «Дружній ШІ» (Е. Юдковський).

Лекція 9. «Сумісні з людиною» ЗШІ.

1) Агентичність (agenticness) – міра та межа автономії інтелектуального агенту.
2) Функціонально обмежені Інтелектуальні агенти, розумніші за нас: чотири випадки.
2.1. ШІ-Суверен.
2.2. ШІ-Оракул.

2.3. ШІ-Джин.
2.4. ШІ-Інструмент.
2.5. Генеративний ШІ: реалії та перспективи.
3) (Додатково) Великі мовні моделі як інструментальний прототип ШІ-Оракула.

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 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 чиїх інтересах діяти: намагатись врятувати свого пасажира 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-пасажира, котрий, обираючи стратегію «рятувати мене» фактично інструктує машину «не рахуватися з життями інших» на свою користь? Але ж чи не є зворотне 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 ШІ-громадян на людське суспільство та потенціальні зсуви у сфері владних структур. Наскільки такі занепокоєння, і які саме, Ви вважаєте обґрунтованими? Якщо слід відмовити ШІ в такому праві, аргументуйте, чому саме.

Семінар 2.
Тема: ШІ сьогодні та ШІ майбутнього.

Завдання 1. Припустимо, що проблеми доступу, змісту та контексту, притаманні сьогодні ШІ-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 ШІ-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» 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?

Семінар 3.
Тема: AGI та його виклики.

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 2. Ви – лідер команди, яка займається розробкою дуже просунутої системи ШІ, 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 ШІ 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?

Семінар 4.
Тема: Екзистенційні ризики та суперінтелект.

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. Розгляньте гіпотетичне майбутнє, в якому люди створили AGI і він не вийшов з-під контролю і не став екзистенційною загрозою для людства. Тоді, мабуть, ШІ 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 ШІ безумовно і перш за все допоможе нам впоратися (яким чином він нам допоможе – відповідати необов’язково, лише за бажанням, на Ваш розсуд).

Семінар 5.

Тема: Принципи індивідуації та проблема контролю.

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 ШІ? Що, як в ході власного розвитку, ШІ зможе обійти наші обмеження на наявність мети? Яким може стати відношення ШІ до людей, якщо ШІ збагне, що саме його розробники є причиною початкової відсутності в нього мети? Яким чином можна запобігти в такому випадку війні людства та машин7

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

Завдання 2. Ви – суддя у справі, де корпорація створила ШІ 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 ШІ-персону, лише як корпоративну власність, на підставі чого вимагає надати їй право знищити ШІ, якщо він перестане слугувати цілям корпорації. Яке саме рішення і на чию користь приймете Ви у цій справі? Які фактори вплинуть на рішення і якого роду співвідношення прав Ви встановите у справі ШІ проти корпорації-розробника?

Семінар 6.
Тема: Автономність інтелектуальних агентів.

Завдання 1. Уявіть, що існує ШІ, 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 2. Уявіть, що ви є частиною міжнародного комітету-робочої групи, перед якою стоїть завдання розробки етичних регулятивів та інструкцій стосовно використання ШІ 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*?

(* до прикладу, заборону використовувати її в міській забудівлі чи навпаки, регулятив використовувати її *тільки* у специфічних зазначених місцях (із

переліком таких місць))

Семінар 7.
Тема: Соціально етичні колізії в контексті Загального ШІ.

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 ШІ, 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 ШІ-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 ШІ-персону буде позбавлено всіх громадянських прав, якщо буде достеменно відомо, що цілі ШІ-персони конфліктують з інтересами людей; позбавлені прав ШІ 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?

Семінар 8.
Тема: Етичні аспекти реалізації AGI.

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 преференціями. 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?

(* Подумайте про місто *поганих*, злих людей, для яких пріоритетом є навмисне нищення довколишнього середовища або заподіяння максимальної шкоди одне одному непрямим способом

– які вказівки з витончених форм садизму можуть дати вони своєму ШІ; уявіть Дім Харконненів із всесвіту «Дюни» Френка Герберта, тільки «powered by AI edition»).

Завдання 2. 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 відредагувати алгоритми передбачення та вибору в ШІ, аби уникнути дискримінації?

Семінар 9.
Тема: Інтелектуальні агенти «сумісні з людиною».

Завдання 1. Яка з чотирьох можливих версій реалізації ШІ здається для Вас найкращою чи найбільш оптимальною: ШІ-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 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?

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 домашньої контрольної роботи (теми ДКР пропонуються на другій лекції; ДКР має бути виконана і здана 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 силлабусу.

Політика та контроль

7.Політика навчальної дисципліни (освітнього компонента)

Відвідування занять

Відвідування лекцій, семінарських занять, а також відсутність на них, не оцінюється. Однак, студентам рекомендується відвідувати заняття, оскільки на них викладається теоретичний матеріал та розвиваються навички, необхідні для отримання певних позитивних результатів вивчення дисципліни.

Вагома частина рейтингу студента формується через активну участь у заходах на семінарських заняттях. Система оцінювання орієнтована на отримання балів за активність студента, а також виконання завдань, які здатні розвинути практичні уміння та навички. Тому пропуск семінарського заняття не дає можливість отримати студенту бали у семестровий рейтинг.

Пропущені контрольні заходи

Якщо контрольні заходи пропущені з поважних причин (хвороба або вагомі життєві обставини), студенту надається можливість виконати контрольне завдання протягом найближчого тижня. В разі порушення термінів і невиконання завдання з неповажних причин, студент не допускається до складання заліку в основну сесію.

Повторне написання контрольної роботи не допускається.

Академічна доброчесність

Політика та принципи академічної доброчесності визначені у розділі 3 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: https://kpi.ua/code.

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	72	8	9	72
2	МКР (2 год.) (МКР може складатися з двох частин по 1 год. кожна)	28	28 (14 х 2)	1 (2)	28
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 – 8. Максимальна кількість балів на всіх семінарських заняттях дорівнює 8 балів × 9 видів робіт = 72 **бали.**

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.

- Відповіді на семінарських заняттях. Ваговий бал – 12. Кількість відповідей - **5**. (Максимальна кількість балів - 60).
- Участь у обговоренні питань семінарів. Ваговий бал – 4. Кількість відповідей - **3**. (Максимальна кількість балів - 12).

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 — **8** балів;

“добре” – достатньо повна відповідь (не менше 75% потрібної інформації) або повна відповідь з несуттєвими недоліками, які допускає студент –**7-6** балів;

“задовільно” – неповна відповідь (не менше 60% потрібної інформації), студент засвоїв основний теоретичний матеріал, але допускає неточності **-5 - 4** бали;

“незадовільно”— відповідь не відповідає вимогам до «задовільно» – **3-0** балів.

2. Виконання модульної контрольної роботи:

Ваговий бал **- 14**. Загальна кількість МКР протягом курсу: 2. Максимальна кількість балів – 28.

14 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.

10-13 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.

7-9 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.

0-6 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	Друга атестація

	7-8 тиждень семестру	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).

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 30 балів. Другою частиною є – залікова співбесіда за трьома питаннями залікового білету, котрий здобувач тягне на початку заліку.

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 ДКР / 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

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.

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 (Telegram, google документи).

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 № 15 від 29.01.2024 р.)
Погоджено Методичною радою КПІ ім. Ігоря Сікорського (протокол № 5 від 29.02.2024 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 зо 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 імплементації 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

І. 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.

ІІ. Теми 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 вироки, 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 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.