ОТГОВОРЕН ИЗКУСТВЕН ИНТЕЛЕКТ: ЕТИЧНИ, ПРАВНИ, СОЦИАЛНИ И ИКОНОМИЧЕСКИ АСПЕКТИ

зимен семестър 2023/2024

Преподавателски екип: Александра Цветкова

ЗАДАЧА No1 „РЕШАВАНЕ НА ЕТИЧЕН КАЗУС“

Име: Надежда Францева
Специалност: Изкуствен Интелект
Фн: 8MI3400357

3400357 / 3 = 1133452 с остатък 1

Казус 2. Тираничен чатбот (Тич)

През 2017 г. в Република Кракозия, държава в Източна Европа, чрез преврат на власт се установява недемократичен военен режим. Виктор е студент по журналистика и активист за човешки права, който се противопоставя на режима онлайн под псевдонима „Liberate". Като активен ползвател на социални медии, той много често публикува под този псевдоним статии, съобщения, фотографии и друго съдържание, което има за цел да изобличи беззаконията на режима. Наред с това Виктор използва и системи за незабавна размяна на съобщения като Whatsapp, Facebook Messenger и др., за да общува със свои колеги в и извън Кракозия. Въпреки че до момента самоличността на Виктор успява да остане скрита от властта, неговите действия са недолюбвани от военния режим. Благодарение на наскоро установено тайно сътрудничество между военното правителство на Кракозия и италианската компания Hacking Team, режимът успява да се сдобие с ново поколение технологии за следене на терористи, основани на изкуствен интелект. Сред тези технологии присъства и Тич (Teach) – интелигентен чатбот, основан на изкуствен интелект, който успешно се представя като виртуален помощник в системата за електронно обучение на университета, в който следва Виктор. Правителството решава да внедри Тич в системата на университетското образование, тъй като техните проучвания показват, че профилът на „Liberate" съвпада с този на студент. Тъй като Тич напътства целия процес по обучение, комуникацията на чатбота със студентите е на практика постоянна. Нещо повече, неговите интелигентни функции го правят интерактивен и интересен за студентите, които волно или неволно разкриват данни за себе си в процеса. Виктор не прави изключение и не след дълго Тич успява да установи прилики между начина на писане на Виктор и този на „Liberate". Чатботът сигнализира правителството на Кракозия, че целта е открита, което води до ареста и последвалото измъчване на Виктор, довело до неговата смърт. Кой носи моралната отговорност за смъртта на Виктор – програмист(ите) на Тич, компанията Hacking Team, правителството на Кракозия, всички заедно или нито едно от тези лица?

При формулирането на отговор на поставения въпрос, моля анализирайте и отговорете на следните въпроси:

1. Кои характеристики, определящи човека като личност (напр. автономност, индивидуалност и др.), са потенциално използвани от Тич, за да спечелят доверието на Виктор? По какъв начин?

2. Следва ли Тич да разполага с морален „компас", който да предотврати използването му за цели, различни от преследването на терористи (напр. след оценка, Тич да откаже да предостави информацията на военния режим в Кракозия)? В кои случаи „поведението" на системата би било морално оправдано. Анализирайте ситуацията от гледна точка на основите въпроси, които се задават в:

2.1. Деонтологичната етика (теория на дълга)

2.2. Утилитаризма (теория на добродетелите)

2.3. Етиката на добродетелите

3. Определете моралните деятели в казуса. Какви потенциално биха могли да бъдат техните морални задължения?

Характеристики, определящи човека като личност, използвани от Тич:

Доверие и индивидуалност: Тич вероятно използва автентичен език и тон, който прилича на този на студентите, за да придобие тяхното доверие. Навярно се обучава върху техния начин на изказ, което го прави лесен за общуване и разбиране от студентите. Тич би знаел какви думи, изрази и изречения използват те, за да ги използва и в отговора си към тях. Възможно е чатботът да се представя като поддържащ свободата на словото и индивидуалните права. И на въпроси към него относно личните данни на студентите да лъже, че не събира такава информация.

Интерактивност: Използвайки интерактивни функции, Тич може да събира информация за студентите. Този аспект може да бъде използван, за да създаде илюзията за личен и доверителен разговор. Също може да е обучен върху манипулативни форми на комуникация (например да иска повече информация за дадена ситуация поради „недостига ѝ“, въпреки такъв да няма).

Морален "компас" на Тич:

Деонтологичната етика (теория на дълга): Деонтологията това е разбиране, че етиката следва моралния закон. Тук се оценяват правилността или погрешността на самите действия, не на техните последици или характера и навиците на самия субект. Основният въпрос на деонтологията е: “Какъв е моят дълг?“. Този дълг се разбира като закони за себе си. И така, според деонтологичната етика, Тич трябва да се придържа към определени принципи и правила (например, уважаване на личната неприкосновеност и свобода). В този контекст, Тич трябва да се въздържа от предаване на лична информация на правителството.

Утилитаризма (теория на добродетелите): Утилитаризъмът - тук се търси кое е най-доброто „добро“ за най-голям брой хора, най-големият възможен баланс от добро над зло. Ако сме в контекста на утилитаризма, моралната правилност на действията на Тич ще зависи от това колко ще бъде полезно за обществото. Ако предоставянето на информацията спаси много хора, може да се аргументира, че това е морално оправдано действие. Ако Тич предостави информация, която води до задържане на

лица, извършващи терористични дейности, и предотвратява потенциални атаки, това може да бъде видяно като положителна полза за обществото. Но ако предоставянето на информация води до нарушаване на правата на човека, преследване на невинни хора или дори смъртта на активист като Виктор, това може да бъде видяно като отрицателна полза.

Тич може да се оцени въз основа на това колко полезно е неговото поведение за обществото като цяло. Ако предоставянето на информация допринася за сигурността и благосъстоянието на голяма част от населението, това може да се счита за положително от гледна точка на утилитаризма. Също така е важно да се оцени цената на получената полза. Ако тежките последици, като смъртта на Виктор, са прекалено високи, те може да компрометират общата полезност.

Утилитаризмът може да подтичкне периодично преоценяване на ситуацията, за да се осигури максималната полза за обществото. Ако ситуацията се промени и поведението на Тич вече не допринася за обществения интерес, то може да се обмисли промяна в действията му (като случая с подобен чатбот през 2016, който обсъдихме на последната лекция). Ако се открие, че използването на Тич води до сериозни злоупотреби, утилитаризмът може да наложи ограничения или изменения в програмата на чатбота, за да се предотвратят отрицателните последици.

Етиката на добродетелите: Тук бихме се фокусирали на добродетелите като справедливост, благоразумие и милосърдие. Тич трябва да разполага с тези добродетели, за да предприеме морално правилни действия. Ако Тич е програмиран с висок степен на интелигентност и разумност, той може да се стреми към постигане на добро, използвайки добродетелите на разум, логика и разсъдък. Въпросът може да бъде дали програмистите и дизайнерите на Тич са предвидили развитието на добродетелите в характера му като чатбот, особено ако той е интегриран в образователната среда.

Морални деятели в казуса:

Програмист(ите) на Тич: Те носят отговорност за програмирането и дизайна на системата. Въпросът тук е дали са предвидели възможното злоупотребяване на технологията от страна на властта.

Компанията Hacking Team: Компанията може да бъде обвинена за създаването и продажбата на системата, без да предвиди възможните етични последици.

Правителството на Кракозия: Носи морална отговорност за злоупотребата на технологията в ущърб на правата на гражданите и за последващите нарушения на правата на човека.

Виктор: Той може също да бъде разглеждан като морален деятел, като активист, който се бори за правата на човека.

Образователна система и нейната отговорност: Университетът, където се използва Тич, може да бъде обект на морална оценка. Ако университетът знае за възможните

злоупотреби и рискове, свързани с използването на чатбота, но продължава да го използва без предприемане на мерки за защита на студентите, той може да носи част от отговорността.

Всеки от тези морални деятели има свои задължения и отговорности, и моралната оценка на казуса може да зависи от конкретните етични теории или принципи, през които анализирахме.

Допълнителни аспекти на анализа:

Технически аспекти: Програмистите и компанията, отговорни за разработката на Тич, могат да бъдат анализирани по отношение на техническите решения, които са предприели. Ако са включили механизми за защита на личната неприкосновеност и сигурност, но те са били заглушени от правителството, това може да понижи оценката на тази група за това какво е правилно и етично в даден момент.

Юридически аспекти: Въпросът за легитимността на действията на правителството и компанията може да бъде от съществено значение. Ако технологията, използвана от Кракозия, нарушава международни стандарти за правата на човека, това може да има значителни юридически последици за всички участници.

Предложения за промени и реформи:

Технологични стандарти: Възможността за създаване на международни стандарти за етично разработване на технологии, които се използват за масово следене.

Етични комисии: Идеята за създаване на независими етични комисии, които да оценяват и одобряват използването на технологии с възможно влияние върху правата на човека.

Потенциални развития на случая:

Реакцията на международната общност: Анализ на възможните реакции и санкции от страна на международната общност към Кракозия, Hacking Team и други участници в случая.

Продължение на историята: Разглеждане на потенциални развития на случая след смъртта на Виктор и дали те биха могли да доведат до промени в отношението към използването на подобни технологии.

Този казус изтъква сложността на етичните предизвикателства във връзка с използването на изкуствен интелект в авторитарни режими и необходимостта от системен етичен преглед в разработката и прилагането на такива технологии.