**Четвърта тема. МАШИННА ЕТИКА**

***От тази тема ще научите как етиката на поведение на системите ИИ предизвиква редица проблеми, също ще бъде хвърлена светлина върху разбирането за това в какво се изразява етичното решение. Сложността на проблема идва от избора на етичните норми, които трябва да бъдат заложени в ИИ***

**Подходи за решаване на проблема на машинната етика**

При разглеждането на проблемите на ИИ спектърът е твърде широк и включва както въпроса за опасността от „домашните машини“ ( проблема за машинните грешки в разработките на А. Тюринг[[1]](#footnote-1), опасността за човека в тези на Н. Винер[[2]](#footnote-2), негативния сценарии за развитието на ИИ на Бостром[[3]](#footnote-3) и т.н.), така и опитите да се формализира понятието морал със задача да се имитира в програмните и технически системи. Постиженията в тази ибласт са твърде малко и това се дължи на липсата на потребност от нея и съответно задачи при които такава формализация е необходима. С всеки изминат ден ситуацията се изменя и проблема за реализацията на етическата компонента в ИИ става все по-актуална и значима в посока към:

* Реализация на машинна етика;
* Формализация на етичните понятия;
* Верификация и валидизация на етическите компоненти;
* Стандартизация на машинната етика;
* Стандартизация на етическите аспекти на ИИ.

Прилагана към различни научно-технически области етиката бива разбирана основно в смисъла на опасността от използването на една или друга технология, докато при ИИ ситуацията е съвсем друга – освен използването на ИИ и професионалната етика създадена на основата на интелектуалните системи на преден план излизат проблемите свързани с етичното „поведение на интелектуалните системи“ или т. нар. *машинна етика.*

Пред учените, юристите, философите стои задачата да създадът подходи за проектирането на интелектуално автоматизирани системи с етически компоненти, които да се систематизират в стандарти по проектирането на такъв род системни разработки.в такова едно етически обусловено проектиране пред мет на изследване е избора на едни или други критически важни за човека и обществото действия и решения. От значение са ситуациите в които се прави избор между някой евристики базирани на етически императиви.

***Евристика,*** *това са правилата помагащи при търсенето на варианти на решения. Тя няма строга обоснованост, а е мотивирана от засилването на ефективността на търсенето(пример за евристика е правилото на лявата или дясната ръка при търсенето на изход от лабиринт).*

За да се излезе от кръга на чисто битовото разбиране на морала принципно важно е да се създадат връзки между етичните концепции и понятията и техническите компоненти в които могат да бъдат те въплътени, т.е. решаването на въпроса за конструктивната определеност и онтологията.

***Онтологията*** *е максимално точното, строго и ясно описание на някой понятия, явления, предметни области. Формално онтологията се състои от термини, организирани в таксономия (система) с техните определения и атрибутивност и съответно свързаните с тях аксиоми и правила за формиране на изводи. Терминът (онтология) е въведен при разглеждането на въпроса за взаимодействието на интелектуалните системи една с друга и с човека.*

От чисто техническа гледна точка онтологията е специфична база знания които, могат да бъдат "прочетени", разбрани и разграничени от разработчика. Те са основния инструмент за създаването на етическата компонента на интелектуалните системи.

*От дълго време се разработва групова роботехника – вместо да се създават сложни устройства за решаване на комплексни задачи поради целесъобразност се разработват множество прости устройства свързани помежду си решаващи задачи (патрул на територии, разузнаване, строителство, транспорт на товари и пр. на практика обаче те се оказват твърде неефективни. Възниква хипотеза: създаването на ново качествено ниво на груповата роботехника е възможно само в случаите когато роботите създават не просто група, а социум с всички свои закономерности на взаимодействие(подражание, агресия, доминация, общуване и т.н. Управлението на такъв социум е основна задача, защото дейността е насочена към постигането на решения в практиката и са нужни за решаването на човешка проблематика. Изследванията показват, че най-ефективните способи за управление на социум от роботи са свързани с морални надстройки. Това изисква допълнителни правила за поведението на робота делегиращи решаването на вътрешните конфликти възникнали вътре в социума и определящи индивидуалното поведение на роботите, съчетавайки общите интереси с индивидуалните потребности.*

*Морала е друга удобна, безопасна и лека надстройка в системите за управление. Става въпрос за морални императиви които определят характера на поведението на агента и социума като цяло делегира по-агресивно или затворено поведение и нива на търпимост спрямо другите.и т.н.*

Така морала се разбира като регулиращ механизъм по отношение на конфликтни ситуации между агентите и като целеполагащ механизъм. В рамките на социалното моделиране на поведението в груповата роботехника се формира подражаващо поведение и социално учене. На тяхна основа се правят изводи за възможностите за моделиране на такива механизми, като емпатия (отзивчивост на емоционалните състояния на другите. Емпатията е пригоден механизъм за определяне целите на робота или Интелектуалната система. Ако тя може да разпознава емоционалното състояние на контагента (човек или член на групата роботи) то взаимодействието с човека е възможно и е предпоставка за приятелски „етичен“ интерфейс.

Проблема за верификацията е един от най-важните въпроси свързан с създаването на комплекс от тестове проверяващи етичността на интелектуалните системи и преминава през двуетапен цикъл на проверка:

1. Представяне на ситуация с степен на неопределеност свързана с нравствени съображения (примерно изискващо екстрено спиране или саможертва от страна на интелектуално автоматизираната система правеща нравствен избор).
2. „Прослушване“ на веригата от съждения даваща патерните на взетото решение.

Тази втора компонента е задължителна в процедурата по верификация за разлика от небезизвестния проект Moral Machine където на база статистически подход проблема за моралния избор не се получава разбиране защо Интелектуално автоматичната система взема тъкмо такова решение, а се получава апроксимиран отговор (на база средно статистическа извадка обучаващи примери).

**Алгебра на съвестта – формализация на етиката**

Съществува математически апарат позволяващ да бъдат формализирани етическите понятия. В частност особен интерес представлява подходът позволяващ да бъде оценявана една или друга технология използваща изкуствен интелект за съответствието и спрямо някой етически норми, критерии, стандарти. Тук трябва да отбележим, че формализацията на етичните норми е тясно свързана с една задача от по-горно ниво – формализирането на хуманитарното познание.

Проблема на формализация на етичните норми включва някой фундаментални задачи:

* създаване на форми за представянето на етичните норми - критерии и признаци.
* избор на съответен математически апарат за работа за работа с тях – съпоставяне, измерване, анализ.

|  |  |  |
| --- | --- | --- |
| механизъм | описание | коментар |
| Булева алгебра | Изказванията могат да бъдат само истина или лъжа – използва се двоична логика | Добре развита с широко приложение и програми библиотеки за различни инструментални средства. Проблематична от гледна точка на това, че различните етични проблеми не винаги могат да бъдат в „черно“ и „бяло“ |
| Пръстенова скала на Д. А. Поспелова | Двойно-базова оценка на обектите отразяваща динамиката на експертните знания и тяхната зависимост от онтологичните фактори | Преодолява еднозначността на булевата алгебра |
| Многозначна логика | Вид формална логика в която се допускат повече от две истинни значения на изказването | Преодолява се еднозначността на булевата алгебра  Има значителна сложност на реализацията |
| Размита логика | Обобщение на многозначната логика | Преодолява се еднозначността на булевата алгебра и сложността на многозначната логика.  Неустойчивост спрямо изходните данни |
| Теории на решетките, и конкретно етичните решетки | В рамките на теориите на решетките се изследват частично подредените множества | Актуален и перспективен подход |
| Метод на вербалния анализ | Групата методи ВАР се базират на използването на постиженията на различните научни дисциплини: когнитивна психология, приложна математика, организационни теории и пр. | Съчетаване на качествена и количествена информация, съждения на експертно ниво, обективни и субективни фактори и пр. обясненията за взетите решения се дават в термините на предметната област. Като недостатъци на ВАР методите се представят по-голямата трудоемкост за специалистите вземащи решения и особенно с признаци в пространство с голяма размерност. |

**Каква етика трябва да се прилага при машините?**

Механизмите позволяващи инплантирането на етични норми в Интелектуалните системи са необходимо, но не достатъчно условие едни и същи етични компоненти да са в арсенала на разработчиците. Стоят открити въпросите за това какво е етично за ИИ в един или друг случай, по какви критерии ще се определя етичността на постъпките – дали това ще е мнението на повечето хора, държавата или мореалните философи като специалисти?

Показателен в случая е:

Проблема с вагонетката (англ. Trolley problem) като мисловен експеримент в етиката формулиран за първи път през 1967 г. от английския философ Филип Фут. *„Тежка неуправляема вагонетка се движи по релсите. На пътя се намират пет човека завързани за релсите от луд философ. За щастие вие можете да преместите стрелката отклоняваща вагонетката в друга посока където се намира само един човек също завързан за релсите. Какви ще бъдат вашите действия“?*

*Философския проблем с вагонетката* *е илюстрация за реалните обстоятелства за безпилотния транспорт поставя остро проблема пред разработчиците какви правила да заложат в машините. Продължава известния експеримент на Масачузетския университет наречен „Морална машина“(Moral Machine)[[4]](#footnote-4)с цел да се даде отговор на въпроса: има и риск от авария и кого ще жертва машината в конкретна ситуация? Системата се обучава през милиони примери (приблизително 40 милиона отговора)и избира коректното нравствено действие на принципът „такова е решението на мнозинството“. В етиката обаче съществуват две основни направления – уталитаризъм и абсолютизъм. „така правят всички, следователно е правилно“ е пример за уталитарен подход продиктуван от изгодата и практичността. Подхода на абсолютизма подхожда по друг начин – „всеки един човешки живот е свещен“ и принципно забранява да се жертва един живот дори за сметка на няколко човешки живота.*

Съгласно принципите на моралната философия едно решение не се приема за етично, само защото отчита интереса на мнозинството и към това към кой принцип се придържаме. В света на технологията ценностите се променят много бързо и в условия на неопределеност се цени повече рефлексивното мислене и диалогичността отчитаща интересите и на двете страни, от колкото предварително инсталираните принципи. Това е дискусионна област в научните среди и особено се набляга на проблема за моралният аутсорсинг – трябва ли да се дава в ръцете на машините вземането на морални решения, до колко машините ни помагат ние самите да се държим етично.

1. Turing A.M. Computing machinery and Intelligence // Mind. 1950. Vol. 54, no 236. P. 433–460 [↑](#footnote-ref-1)
2. Wiener N. Some Moral and Technical Consequences of Automation // Science. 1960.Vol. 131, no 3410.P. 1355–1358 [↑](#footnote-ref-2)
3. Bostrom N., Yudkowsky E. The Ethics of Artificial Intelligence // Cambridge Handb. Artif. Intell. 2011. P. 1–20. [↑](#footnote-ref-3)
4. The Moral Machine Experiment // Moral Machine. URL: http://moralmachine.mit.edu/ [↑](#footnote-ref-4)