Анализ равновесных состояний AI в медицинских системах

предложение

Илья Юхновский

Теорема Лёба

PA-арифметика Пеано

P-любая формула, доказуема в *PA*

Если "*если P доказуемо в PA, то P истинно*" = *true*

тогда "P доказуемо в PA" = *true*

Теорема Гёделя

Формальная арифметика (ФА), формальная система - то где можно определить основные арифметические понятия:

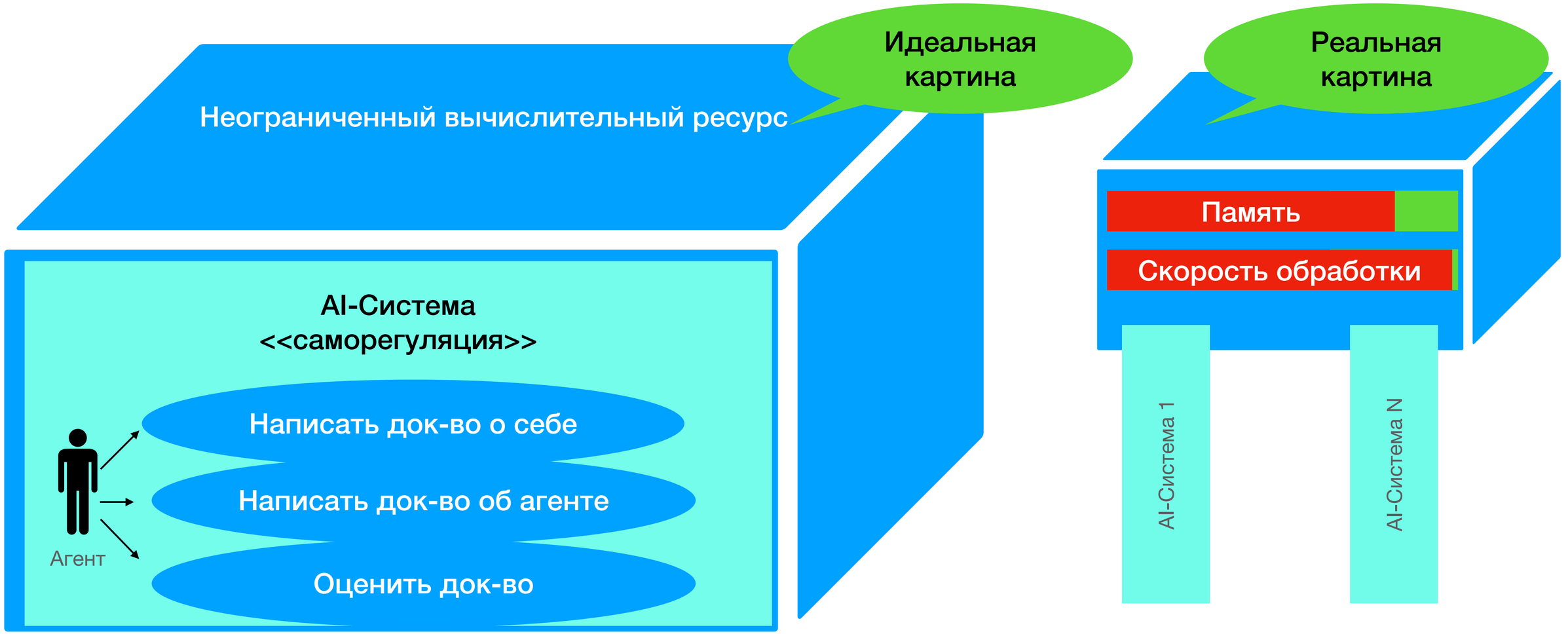
- натуральные числа
- 0
- 1
- сложение
- умножение

Теорема 1 если ФА непротиворечива, то в ней существует невыводимая и неопровержимая формула

Теорема 2 если ФА непротиворечива, то в ней невыводима некоторая формула, содержательно утверждающая непротиворечивость этой арифметики

Модель исследований

Теорема Лёба и теорема Гёделя позволяют делать предсказания о поведении систем, способных к саморегулированию с неограниченными вычислительными ресурсами, с помощью которых можно писать и оценивать доказательства



Вопросы исследования

Как перейти от предсказания поведения теоретической системы к практической - нейронной сети?

Как ресурсные ограничения влияют на теорему Лёба?

Метод исследования

Рассмотреть равновесные состояния системы (нейронной сети):

- классическое равновесие по Нэшу
- коррелированные равновесия - взаимно кооперативное программное равновесие

Агенты системы (слои и узлы нейронной сети) - игроки в теории игр.

Ядро состояний - слабый вариант равновесия Нэша, набор состояний, в каждом из которых ни одна группа акторов, способных выстроить новое (отсутствующее в данном ядре) состояние, не улучшит своей ситуации по сравнению с их состоянием в данном ядре.

Программное равновесие - равновесие "мета-игры", выбора в какую программу играть. Каждый игрок представляет собой алгоритм, который может читать исходный код своего оппонента.

Метод исследования (продолжение)

В равновесных состояниях изменение стратегий:

- написания доказательств о себе
- написания доказательств о другом агенте
- оценка доказательств

не улучшает одновременно устойчивость системы, что характеризует:

- оптимальность алгоритма
- воспроизводимость результатов

а следовательно позволяет произвести оценку достоверности.

Неопределенности

Нейросеть AI

Если мы знаем, какой алгоритм реализует нейросеть, то есть две неопределенности:

1) неопределенность исходных данных - связана с эмпирической неопределенностью, мы должны наблюдать за входными данными и применять теорему Байеса, чтобы определить вероятности результата на выходе

2) неопределенность того, что действительно делает алгоритм - касается логического факта неопределенности алгоритма вычисления. Теория вероятности не решает эту проблему, потому что вероятность не может быть присвоена логическим фактам. (1+1=2)=>Ф, но так как если A=>B, то P(A) <= P(B) отсюда уверенность в 1+1=2 такая же как и в Ф. Решением является ослабление критерия P(A) <= P(B) до тех пор пока не будет доказана импликация. Но это оставляет открытым вопрос о том, как импликация будет доказана и это возвращает нас к поиску принципиального метода управления неопределенностью в отношении логических фактов, когда отношения предполагаются, но не доказаны.

Значимость исследования

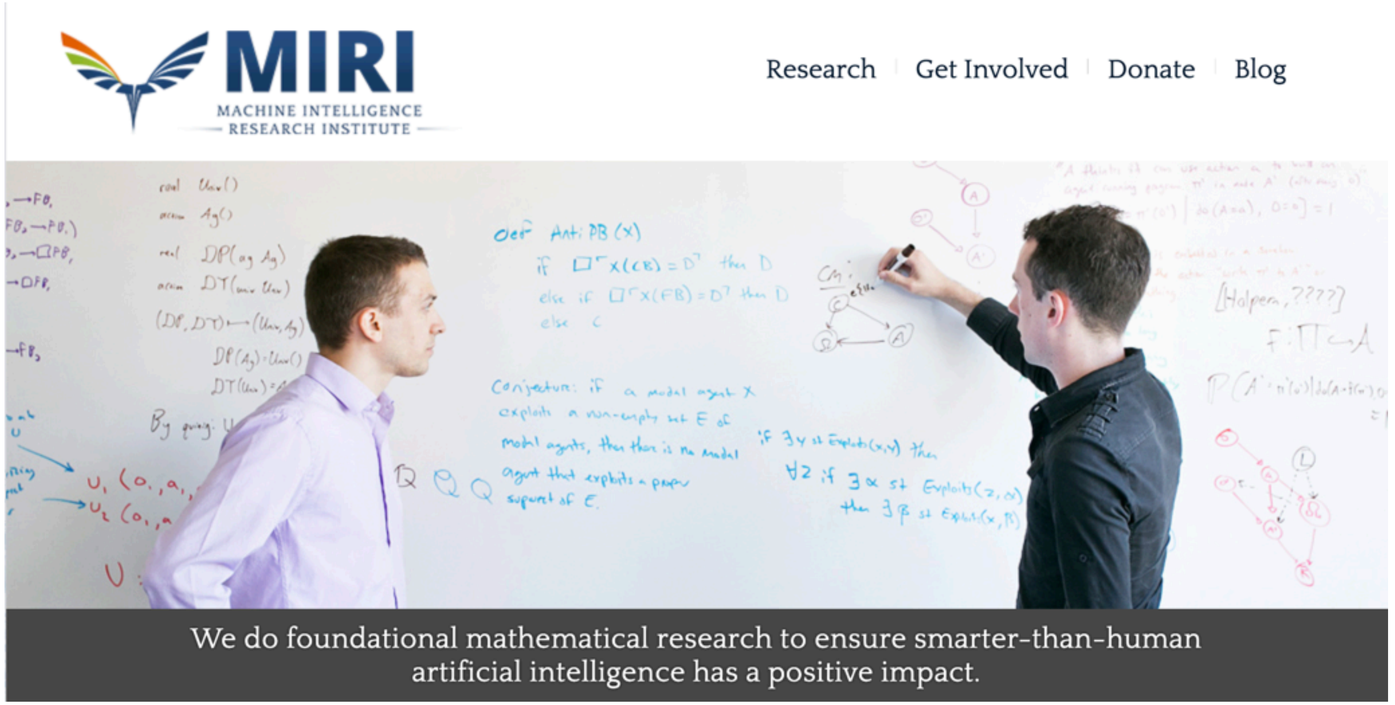
Методика оценки вероятности нахождения агентов системы и
самой системы в целом в ожидаемых (равновесных) состояниях

Зная вероятность - можем можно провести оценку достоверности результатов AI медицинских систем

Конечный результат - сертификация и лицензирование AI в области медицины

Экспертиза

Специалисты по безопасности AI:



НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ им. Р. Е. Алексеева

Кафедра "Биоинженерия и ядерная медицина"