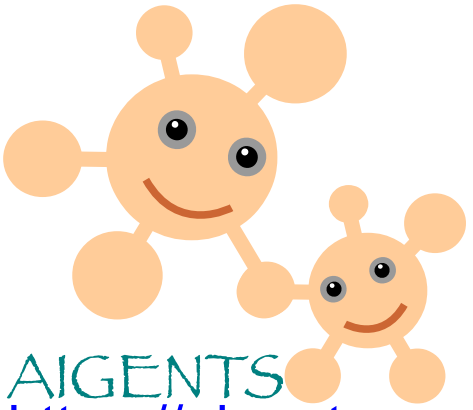
Искусственный интеллект – состояние вопроса и риски

Антон Колонин, к.т.н.
akolonin@aigents.com
anton@singularitynet.io

Новосибирский государственный университет
*НАСТОЯЩАЯ НАУКА



https://aigents.com


SingularityNET
https://singularitynet.io

Где мы находимся?

Программируемый → Адаптивный

Управляемый → Автономный

194?г. 2018г. 20??г.

Слабый → Сильный

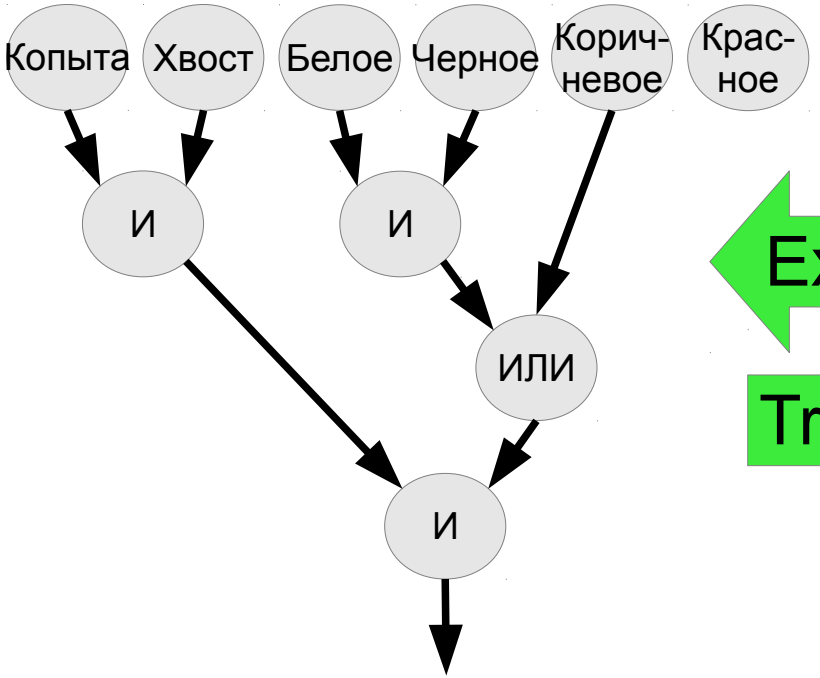
Статистика и машинное обучение → ИИ уровня человека → ИИ сверх-человеческого уровня

Общий ИИ (AGI)

Не взятые рубежи ИИ

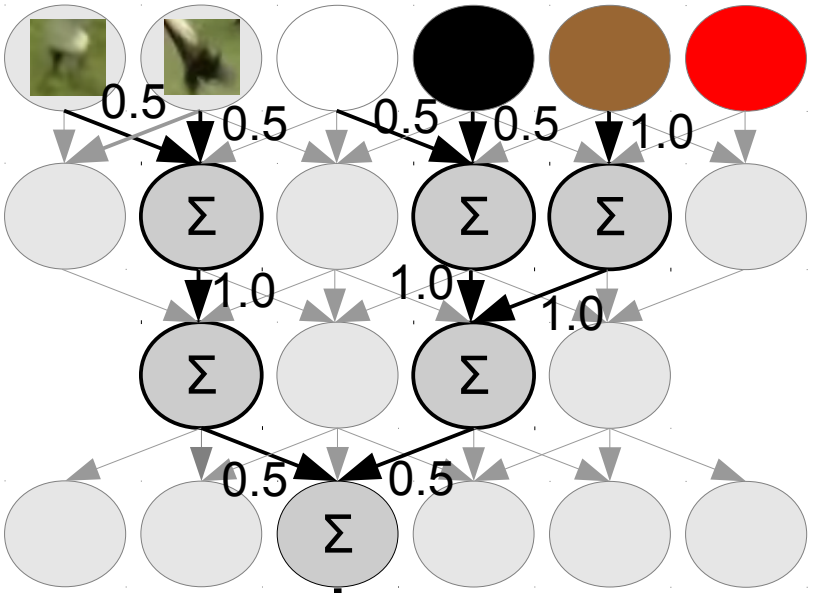
- Объяснимый ИИ (explainable AI)
- Передача знаний (transfer learning)
- Быстрое обучение (few-shot learning)
- Сильная генерализация (strong generalization)
- Генеративные или генерирующие модели обучения (generative models)
- Структурированное обучение и предсказание (structured prediction and learning)
- Решение проблемы катастрофического забывания (fighting catastrophic forgetting)
- Достижение возможности инкрементального обучения (incremental learning)
- Новый тест Тьюринга (например, “Baby Turing Test”)
- Решение проблемы сознания (consciousness)

Интеграция “нейросетевого” подхода и “вероятностной логики” для “transfer learning” и “explainable AI””



Explain

Transfer



(Копыта И Хвост) И
((Белое и Черное) ИЛИ Коричневое)

=> Лошадь

Риски, связанные с ИИ

- Вооружения с автономным высокоскоростным ИИ (LAWS)

  с точки зрения внутренней безопасности;

  с точки зрения внешней безопасности;

  с точки зрения коллективной международной безопасности.
- Автономные адаптивные системы управления

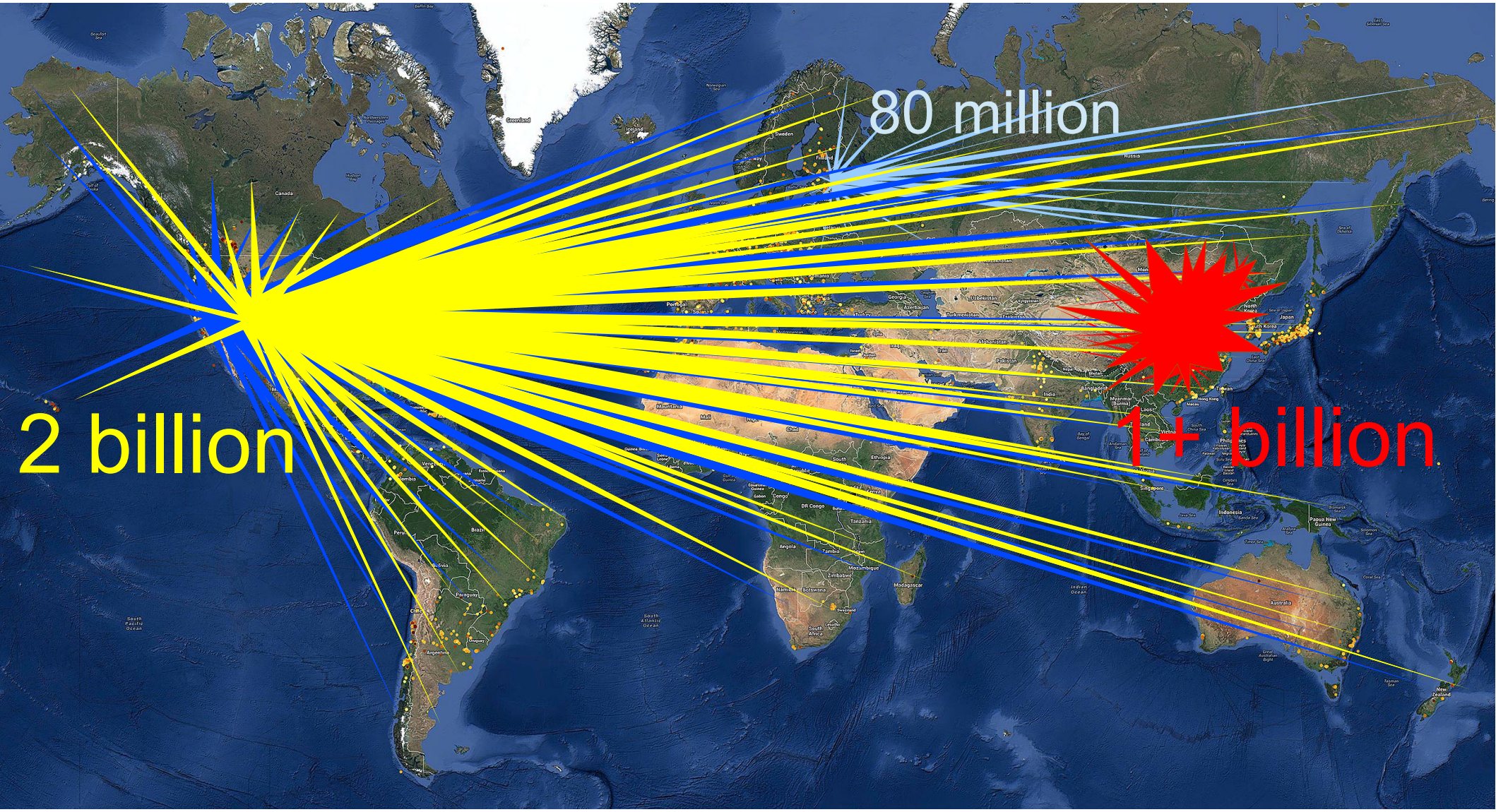
  в том числе – обладающие сознанием или его зачатками.
- Интеллектуальные системы, как средство разделения

  на управляющих ими и управляемых ими (“digital divide”).
- Интеллектуальные системы, как причина деградации

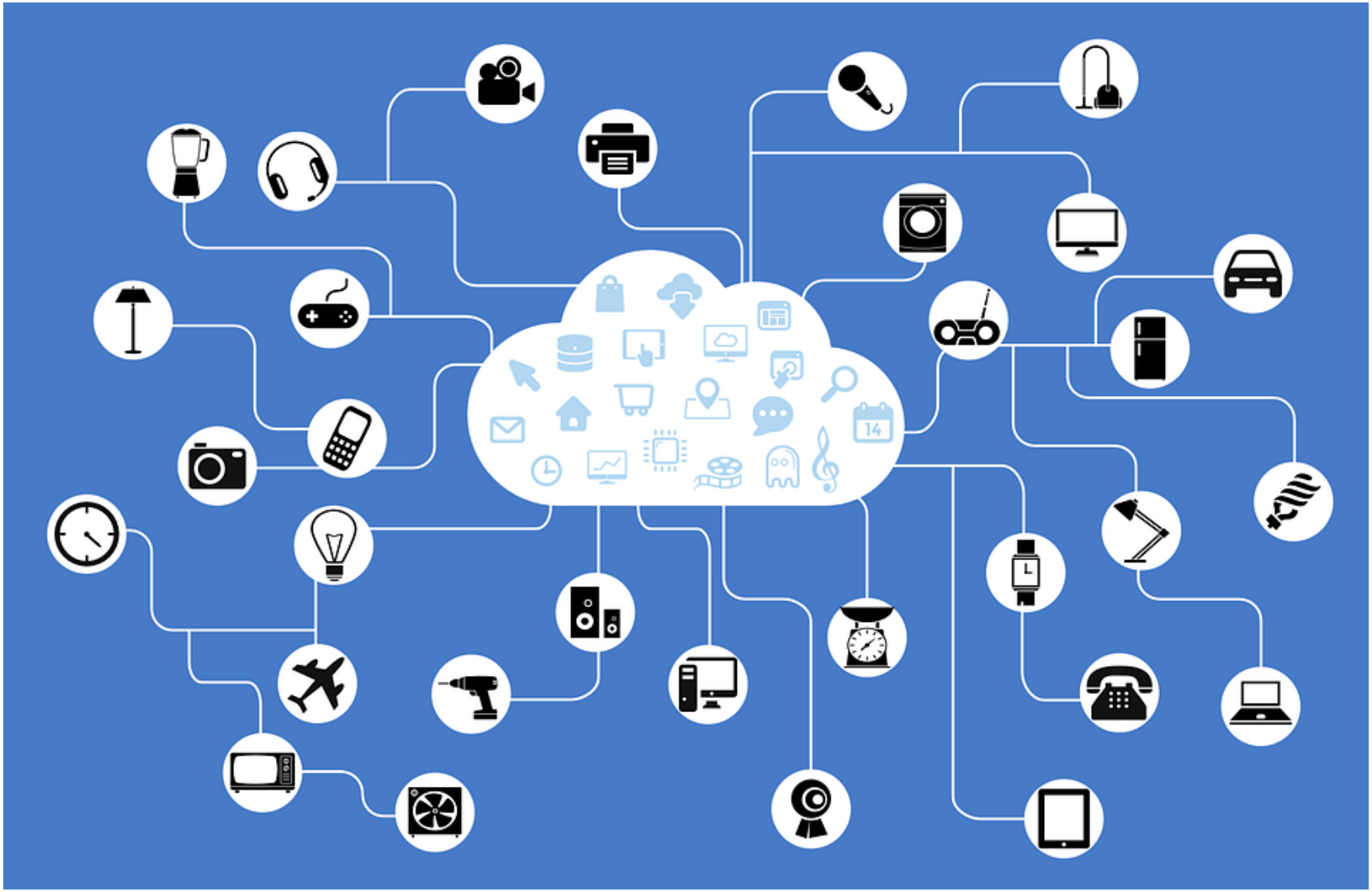
  интеллектуальных способностей человека в перспективе.
- Массовые системы коллективного управления как феномен

  эволюционного масштаба принципиально не управляемый.

Месячная база “сети социальных вычислений”: Google+Facebook – в мире, ВКонтакте – в России WeChat+Baidu+QQ - в Китае



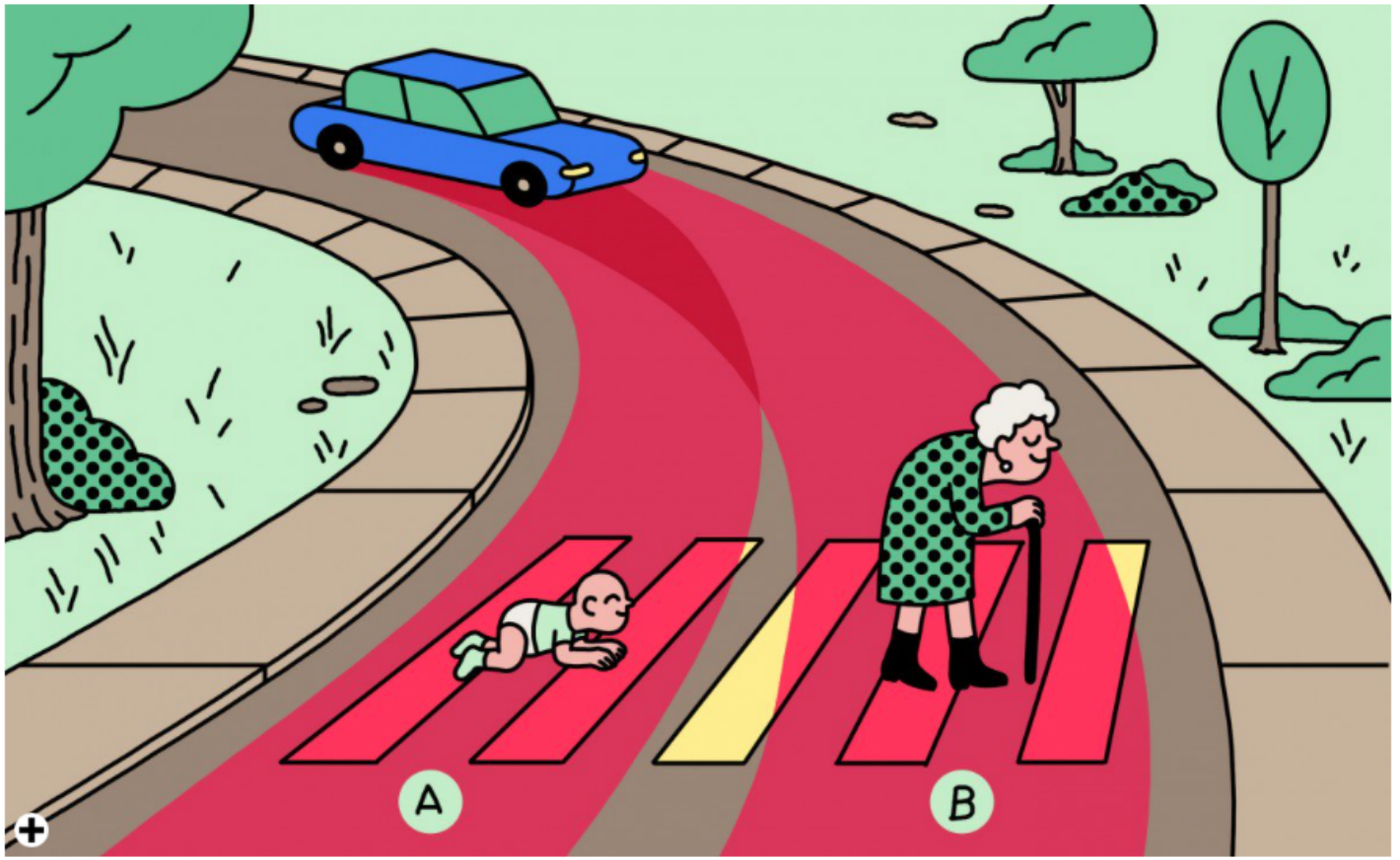
Новая сетевая реальность: 3+ миллиарда людей, пополняется 15 миллиардами устройств IoT в 2018, многие из них дополнятся ИИ в последующем



Меры по снижению рисков

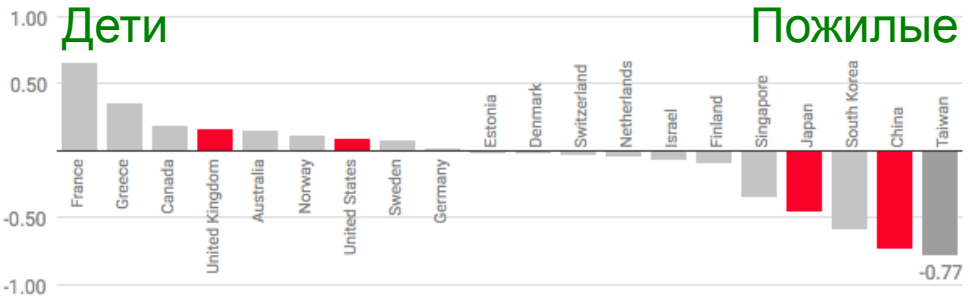
- Международный запрет «автономных систем смертоносных вооружений» (LAWS)
- Решение проблемы «объяснимого искусственного интеллекта»
- Создание технологий для программирования не функций и не целей, а ценностей
- Демократизация доступа к технологиям ИИ для массового пользователя
- Открытость алгоритмов и протоколов систем с ИИ на международном уровне
- Создание систем коллективного интеллекта, устойчивых к манипуляциям и злоупотреблениям

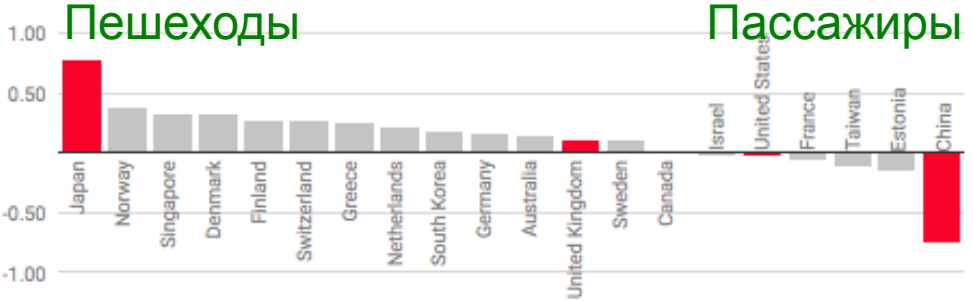
Законы Азимова и этика ИИ



“Проблема Вагонетки”:

Общего решения – НЕТ!

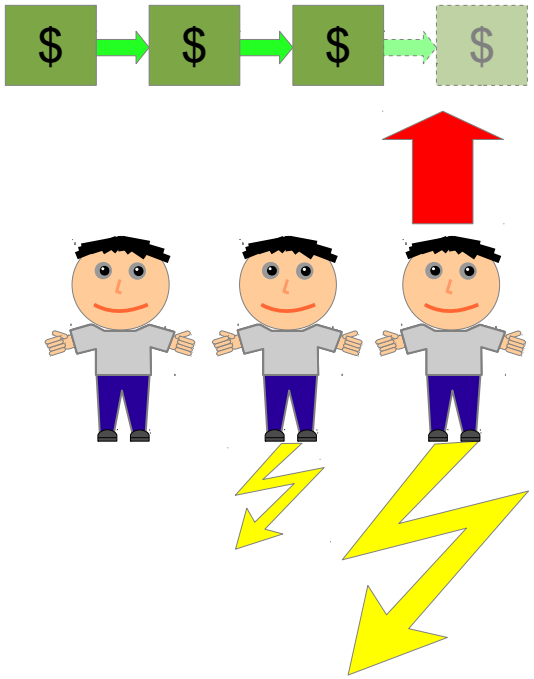




https://www.technologyreview.com/s/612341/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

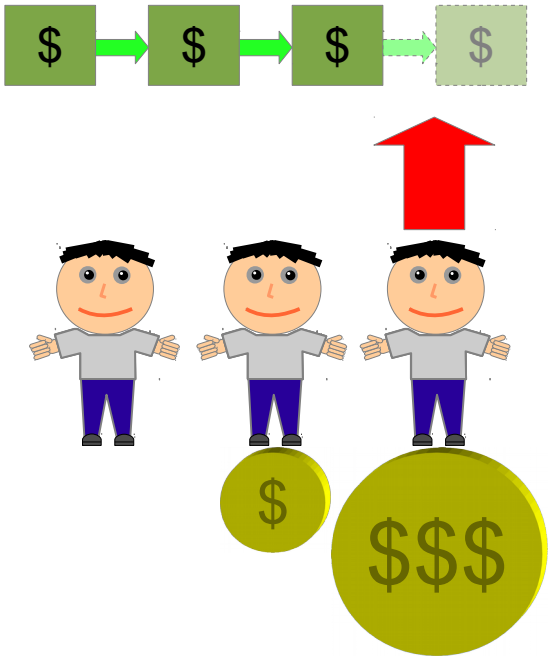
Задача достижения общественного консенсуса в больших распределенных вычислительных и социальных системах

Proof-Of-Work



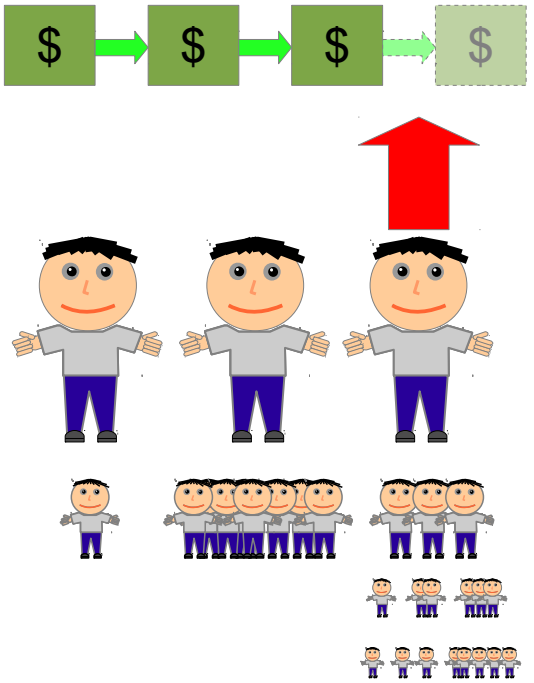
Право силы:
Возможность влиять на принятие решений определяется физическими ресурсами.

Proof-Of-Stake



Право денег:
Возможность влиять на принятие решений определяется финансовыми ресурсами.

Proof-Of-Reputation



$$R_i = \sum_t \sum_j (R_j * V_{ijt})$$

Право авторитета:
Возможность влиять на принятие решений определяется истинной “глубинной” репутацией.

Наш вклад в общее дело:

https://aigents.com

Персональная система искусственного интеллекта для работы в Интернете с конфиденциальностью ваших данных

SingularityNET

https://singularitynet.io

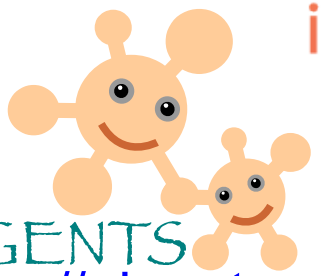
Платформа для построения распределенных систем искусственного интеллекта с открытым кодом и протоколом

Приглашаем партнеров!

Антон Колонин, к.т.н.
akolonin@aigents.com
anton@singularitynet.io

iHeart

zkylos



https://aigents.com

SingularityNET
https://singularitynet.io

TravelChain