Исследовательский семинар
«Этические проблемы искусственного интеллекта»
РАИИ, МГУ, ИФ РАН, 11 декабря 2024

# Человек-оператор и искусственный интеллект: навязанное сосуществование или выгодное партнерство?

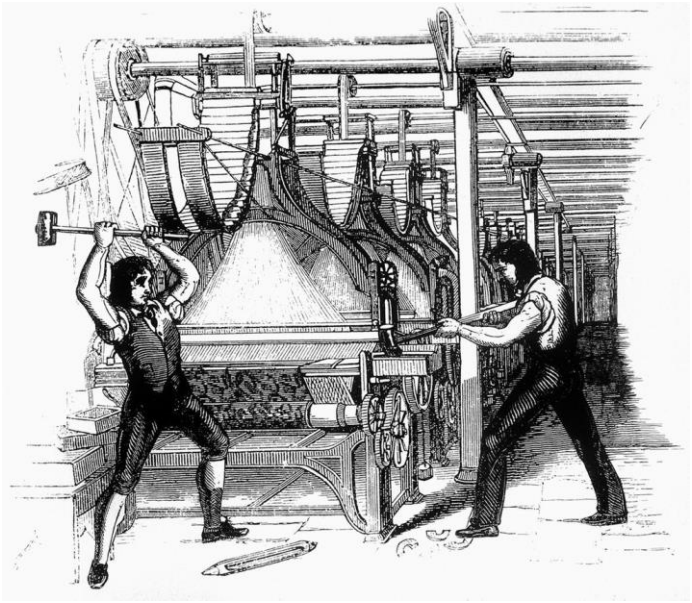
В.М. Дозорцев (Центр цифровых технологий, Rubytech/МФТИ)
А.Л. Венгер (ГУ «Дубна»)

# План рассказа

- Взаимодействие Ч-О и ИИ – краткий исследовательский ландшафт
  - старая проблема доверия оператора технике
  - неготовность использовать ИИ
  - что привносит в проблему появление ИИ
- Как повысить доверие ИИ?
  - традиционные приемы: принцип активного оператора, теория активных систем, оптимальное резервирование Ч-О и ИИ
  - учет индивидуально-психологических факторов доверия ИИ: BigFive, локус контроля, атрибутивный стиль, общее отношение к технике, прошлый опыт
- Модель принятия операторских решений на основе факторов личностной тревожности
  - зависимость предпочитаемых стратегий от уровня тревожности
  - типы стратегий (имитационный эксперимент)
  - устойчивость и подвижность индивидуальных стратегий
- Взаимная адаптация Ч-О и ИИ
  - учет индивидуальных характеристик Ч-О в алгоритмах ИИ
  - специальный тренинг операторов
- Предварительные выводы

# Доверие операторов технике – краткий исследовательский ландшафт

1. Проблема доверия/недоверия (ДНД) технике
   - одна из старейших – луддиты (начало XIX века)
   - проявляется крайне разнообразно – ей занимаются инженерные психологи, социологи, философы, культурологи, писатели-фантасты
   - обостряется на каждом витке усложнения техники



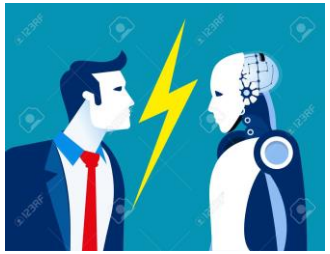
2. Затрагиваются все категории пользователей
   - активные – профессиональные операторы технических систем, эксплуатанты техники вне профессиональной деятельности
   - пассивные – владельцы, клиенты (пассажиры, юзеры)

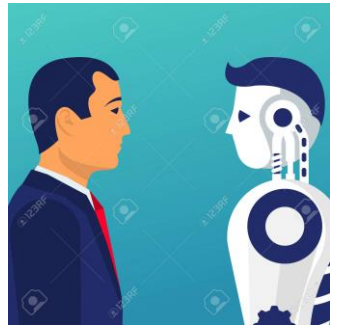
3. ДНД технике = ДНД автоматике
   - задолго до ИИ оператор начал воспринимать объект через компьютеризированную систему управления
   - эффективность работы техники зависит от уровня доверия операторов автоматизированным системам управления (особенно в сложных и неопределённых условиях)

# Доверие операторов технике – краткий исследовательский ландшафт$_2$

## 4. Теоретическая основа

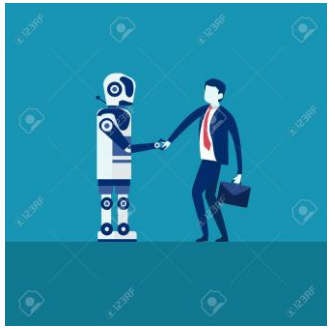
- ДНД рассматривается в контексте использования оператором т.н. агентных систем, в первую очередь – когнитивных агентов (КА)
- ДНД агенту – отношение к тому, поможет ли агент достичь целей оператора в ситуации неопределенности и уязвимости (*Lee, See, Human Factors, 2004, 46*)
- ДНД агенту предполагает восприятие его как обладающего намеренностью (*de Visser, Parasuraman. The social brain. N.-Y., 2011*)





## 5. В отечественной психологии

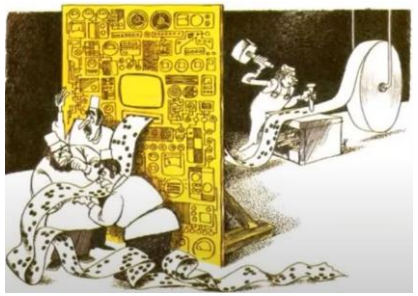
- ДНД – осознанные психологические отношения, описываемые когнитивно, эмоционально и поведенчески (*Акимова, Обознов, Псих. журнал, 2016, 7(6)*)
- по мере усложнения техники ДНД смещается к отношению «субъект-субъект»
- семантизация техники повышает мотивацию человека, формирует когнитивную метасхему объекта, упрощает коммуникацию (*Хорошилов, Вестник НГУ, 2009, 3*)
- пара «доверие-недоверие» несимметрична (*Купрейченко, 2008*)
- доверие – важный операторский ресурс: оптимизирует когнитивную сложность, эмоциональную напряженность, внимание, снижает психосоматику (*Акимова, Обознов, Псих. исследования, 2017, 8*)



# Неготовность использовать ИИ

1. Вечная проблема новой техники
   - Перл-Харбор (7.12.1941) – внезапность атаки, неготовность правильно применить радары
   - сбои в системах противоракетной обороны (70-80 гг.)
   - переход к микропроцессорным системам управления
   - системы управления с прогнозированием

2. Особенности систем ИИ, усугубляющие неготовность
   - неверие в возможности ИИ и технологии вообще
   - неуверенность в способности овладеть инструментами ИИ
   - страх потерять работу
   - проекция – перенос внутренних проблем оператора на ИИ
   - страх уронить престиж – ущемленная гордость хомо сапиенс: ИИ посягает на святая святых (оптимизация, планирование, прогнозирование)
   - оборотная сторона семантизации техники: ИИ воспринимается чужой/соперник/враг




# Что привносит в проблему ДНД искусственный интеллект

1. В промышленной автоматизации мы пока столкнулись с рекомендательными системами (слабый ИИ) - управление в режиме совета оператору

2. Безопасность: промышленная и кибербезопасность, конфиденциальность данных, охрана  окружающей среды, законодательство

3. Прозрачность: лозунг «компьютер сказал нет» больше не проходит!
   - транспарентность решения; объяснения, понятные оператору

4. Коммуникативность
   - чем больше уверенности в намеренности КА, тем больший ресурс внимания оператора необходим
   - сверхдоверие автоматизации (т.н. «уклон автоматизации») – *Dzindolet, 2002, Human Factors, 44*
   - большее доверие КА, подобным машинам (возможно, животным), но не людям
   - большее недоверие злонамеренным КА, похожим на людей, а не на машины (*Zak et al. Hormones and Behavior, 2005, 48*)

5. Мотивация / ответственность
   - мотивационная ловушка: принять правильный совет → «ИИ – умнее оператора»; отклонить правильный совет → «Оператор некомпетентен!»; принять неправильный совет → «А ты куда смотрел?»; отклонить неправильный совет → «Молодец, но мы тебе за это зарплату платим!»
   - кто бенефициар решений ИИ – вендор/разработчик/покупатель/сам ИИ
   - кто отвечает за негативные последствия – никто (форс-мажор), производитель/ разработчик/«учитель»/владелец/пользователь
   - этический кодекс, проблема предвзятости, запрет ИИ в определенных задачах
   - справедливость в получении выгод и доступность ИИ

# Как повысить доверие интеллектуальной технике

1. Современная техника берет все больше операций на себя: быстрое регулирование, программное управление, оптимальное управление и даже предиктивный анализ состояния процессов и оборудования
2. Принцип активного оператора (*Ломов. Человек и техника, 1966*)
   - человек не должен исключаться из цепи управления, сохраняя активные функции
   - его переход к активным действиям проходит с меньшими затратами
3. Теория активных систем (*Бурков и др. Большие системы, 1989*): премии – за принятие верного и отклонение неверного совета; штрафы – за отклонение верного и принятие неверного совета
4. Оптимальное резервирование оператора и автоматики (*Костин, Избранные труды, 2021*)
   - критерии перехода к ручному управлению – качественные; наоборот - количественные
   - гармонизация отношений операторов (склонных не доверять автоматике) и разработчиков (опасающихся человеческих ошибок)
5. Рецепты повышении доверия к ИИ пока небогаты: не допускать ошибок ИИ (иллюзия); готовить Ч-О к возможности и способам устранения ошибок; управлять ожиданиями пользователей; повышать прозрачность и объяснимость ИИ
6. Серьёзный исследовательский пробел: как определенные типы личности преодолевают трудности взаимодействия с ИИ и как ИИ могут адаптироваться к личности пользователя.

Три требования к «доверенному ИИ» (*Riedl, Electron Markets, 2022*):

- регистрация предпочтений и состояний Ч-О
- интеграция этого потока данных в личностный профиль Ч-О
- на этой основе адаптация в реальном времени для повышения доверия оператора

# Учет индивидуально-психологических факторов доверия ИИ

1. BigFive: личностные факторы, характеризующие адаптацию человека к социальной среде
   - экстраверсия, способность к согласию, добросовестность, невротизм, открытость опыту
   - возможно все факторы, кроме невротизма, положительно влияют на доверие
   - невротизм и тревожность могут снижать доверие и сильно влияют на принятие рискованных решений (*Miller et al. , Frontiers in Psychology, 2021*)
   - в рекомендательных системах принять совет ИИ – часто менее опасно, чем отклонить его
2. Атрибутивный стиль оператора (*Селигман. Как научиться оптимизму, 2018*)
   - при пессимистическом стиле доверие ИИ выше
3. Предыдущий опыт работы со сложной техникой
   - чем удачней опыт, тем выше доверие
   - проверка этой гипотезы для случая ИИ в эксперименте
4. Отношение к общим тенденциям развития общества
   - вера в прогресс или консерватизм; доверие институтам или конспирология
   - бессилие рядовых людей и избранность элиты (*Postman. Technopoly. N.-Y., 1993*)
5. Многие факторы маскируются в результате профессионального тренинга, но могут отражаться в невротизме, особенно в ситуациях большого риска
6. Фактор тревожности можно исследовать экспериментально, а затем использовать результаты исследования для адаптации ИИ к индивидуальным особенностям Ч-О

# Исходные положения модели принятия операторских решений в рекомендательных системах

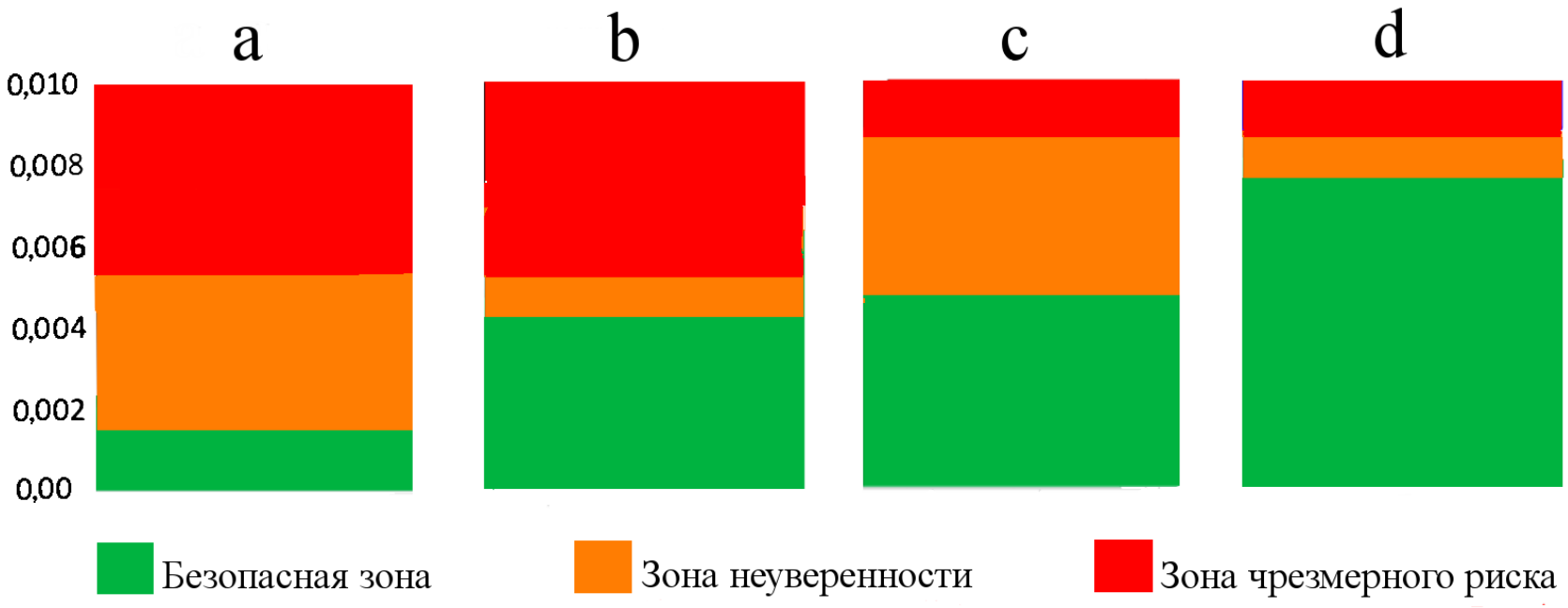
- Теория активных систем (*Бурков и др. «Большие системы», 1989*)
- Математическая теория решений с учетом возможности неприемлемого ущерба
- Теория последовательных решений
- Гипотеза о двух аспектах личностной тревожности (*Венгер, Автоматизация в промышленности, 2013, 7; 2014, 12)*:
  - максимальная субъективно допустимая вероятность катастрофы **(«боязливость»**)
  - субъективно необходимая надежность оценки этой вероятности (**«склонность к сомнениям»**)
- Возможные действия оператора:
  - принять рекомендацию ИИ и остановить процесс (в предположении, что сигнал ИИ правилен)
  - отклонить рекомендацию ИИ и продолжить процесс (в уверенности, что сигнал ИИ – это «ложная тревога»)
  - промежуточное решение: запросить дополнительную информацию и уже на ее основе принимать окончательное решение (вариант небезопасный, т.к. за время получения дополнительной информации может наступить авария)

# Распределение решений в зависимости от субъективной вероятности катастрофы

- Зона принятия риска: согласно субъективной оценке, риск не выше допустимого; субъективная надежность оценки достаточна – несмотря на рекомендацию ИИ принимается решение продолжить процесс
- Зона неуверенности: согласно субъективной оценке, риск не выше допустимого; но субъективная надежность оценки недостаточна – принимается решение о сборе информации
- Зона чрезмерного риска: согласно субъективной оценке, риск выше допустимого – принимается решение об остановке процесса

Компьютерное моделирование показало, что в зависимости от характеристик технологического процесса, точности предсказаний ИИ (частоты «ложных тревог») и величины ущерба оптимальными оказываются разные стратегии (разные соотношения «боязливости» и «склонности к сомнениям»)

# Варианты стратегии оператора




По вертикали – субъективная вероятность катастрофы

a) «боязливый» оператор с высокой склонностью к сомнениям

b) «боязливый» оператор с низкой склонностью к сомнениям

c) «бесстрашный» оператор с высокой склонностью к сомнениям

d) «бесстрашный» оператор с низкой склонностью к сомнениям

# Взаимная адаптация Ч-О и ИИ (три этапа исследования)

1. Выделение и анализ операторских стратегий: $Э_1 \rightarrow П_1 + ИИ_1 + Ч\text{-}О_1$
   - $П_1$ – динамический процесс с помехой; $ИИ_1$ – логический алгоритм со знанием будущего; $Ч\text{-}О_1$ – логический алгоритм с двумя параметрами тревожности
   - большой имитационный эксперимент: в определенных условиях ($П_1$ + $ИИ_1$) любая выраженная стратегия может быть лучше других
2. Выявление и анализ практических стратегий: $Э_2 \rightarrow П_1 + ИИ_1 + Ч\text{-}О_2$
   - $Ч\text{-}О_2$ – студенты профильных вузов, добровольцы
   - гипотеза: сталкиваясь с необходимостью, испытуемые в конкретных условиях ($П_1$ + $ИИ_1$) пытаются (с разным успехом) изменить стратегию принятия/отклонения рекомендаций
3. Специализированный тренинг операторов: $Э_3 \rightarrow П_3 + ИИ_3 + Ч\text{-}О_3$
   - $П_3$ – фундаментальная тренажерная модель ТП (аппарата); $ИИ_3$ – алгоритм предиктивного анализа состояния оборудования с возможностью адаптации к характеристикам операторов; $Ч\text{-}О_3$ – производственники (операторы, технологи, пр. специалисты)
   - цель (1): ИИ адаптируется к оператору – стартуя с априорно определенных характеристик тревожности и постоянно дообучаясь, определяет время и форму информирования об угрозах («боязливость»), состав и детальность дополнительной информации (склонность к сомнениям)
   - цель (2): оператор адаптируется к ИИ в ходе тренинга на высокоточной модели технологического объекта, включающей продвинутый алгоритм ИИ

# Предварительные выводы

1. Искусственному интеллекту в промышленности нет альтернативы
   - опережающее внедрение инструментов ИИ, затрагивающих безопасность людей, активов и инфраструктуры
   - на пороге все более сильный ИИ, проблемы доверия будут обостряться
2. Высокое доверие – предпосылка эффективного взаимодействия Ч-О и ИИ
   - необходимы дальнейшие исследования психологических факторов, определяющих уровень ДНД
3. Взаимная адаптация Ч-О и ИИ – гармонизация элементов человеко-машинной системы в ситуациях неопределенности и уязвимости
   - постоянная перенастройка ИИ на Ч-О
   - специализированный тренинг операторов в присутствии ИИ в системе управления
4. Ч-О и ИИ – не соперники: их преимущества и дефициты практически не пересекаются, а партнерство жизненно необходимо для безопасного и эффективного производства. Такое партнерство невозможно без доверия Ч-О к ИИ, достигаемого в их взаимной адаптации

# Спасибо за внимание!

**Дозорцев Виктор Михайлович** – д-р техн. наук,
директор по развитию бизнеса ООО «Центр цифровых технологий», Rubytech/МФТИ victor.dozortsev@mipt-cdt.ru

**Венгер Александр Леонидович** – д-р психолог. наук,
проф. кафедры психологии Государственного университета «Дубна» venger.1@uni-dubna.ru