НИС «Современные дискуссии в философии науки». 2024-25 гг.

Преподаватели НИС: проф., с.н.с. Е.Н. Князева, доц., н.с. И.А. Карпенко, доц., н.с. В.Э. Терехович, доц. Д.Н. Дроздова.

Цель нашего научно-исследовательского семинара – погрузить студентов в современные дискуссии в области философии науки и способствовать приобретению навыков участия в исследовательской деятельности. Темы отдельных блоков связаны с общими вопросами философии науки, философскими проблемами частных наук, а также отношениями истории науки и истории философии.

Стремительное развитие искусственного интеллекта за последние годы способно существенно повлиять на все сферы человеческой длительности, в т.ч. на науку. Многие важные решения, касающиеся жизни человека, все больше будут зависеть от искусственного интеллекта. Поэтому в этой области так важен гуманитарный анализ с позиций философии науки. **Общая тема НИС 2024/25 года – искусственный интеллект (ИИ)**.

Формы проектной работы
Занятия проводятся в формате мастер-классов преподавателей, обсуждений текстов и докладов студентов, коллективных проектов, дебатов, совместного комментирования текстов. Выбор конкретных форм групповой или индивидуальной проектной работы определяется исследовательскими интересами студентов. За исследовательские проекты, при их надлежащем оформлении, могут быть начислены кредиты.

Студенты смогут выбирать из нескольких форматов домашней письменной работы, благодаря которой получат необходимые навыки для написания мини-статьи по итогам каждого модуля. В дополнительном методическом блоке в 3 модуле (В.Э. Терехович) студенты отработают навыки поиска литературы, написания статей по теме НИСа, работы над КР, подготовке презентаций и выступлений. Студенты, которые пишут КР/ВКР по тематике НИСа, смогут представить их на методических семинарах НИГ «Философия науки», а также в методическом блоке.

НИС «Современные дискуссии в философии науки» организован **НИГ «Философия науки»**. Общий чат НИС и НИГ с анонсами мероприятий. По всем вопросам можно обращаться к секретарю НИГ Д.Е. Лаврищеву (delavrischev@hse.ru, @VTStampede в ТГ).

1 модуль. Рациональность и искусственный интеллект, виртуальные миры, ИИ как инструмент самопознания. Преподаватель: И.А. Карпенко.

В 1 модуле обсудим проблему рациональности искусственного интеллекта: мыслит ли он, если да, то как, о чём может мыслить, каковы условия и возможные следствия его независимого мышления. Рассмотрим следующие вопросы: что такое рациональность; какова основа человеческой рациональности; какие логические системы (инструменты) могут быть применены для описания рациональности ИИ (классическая логика, интуиционистская, релевантная, многозначная, паранепротиворечивая, квантовая логика и т.д.); как соотносится сознание и тело, и может ли помочь нейропсихология в понимании ИИ?

Следующий круг проблем, предлагаемых к обсуждению, связан с разработкой виртуальной реальности и виртуальных миров с помощью ИИ. Рассмотрим вопросы, касающиеся гипотезы симуляции в контексте новых данных: какую роль может играть ИИ при проектировании симуляций, какие миры он может порождать, и мог бы наш мир (гипотетически) быть создан ИИ, и к каким следствиям это приводит.

Последняя группа проблем посвящена ИИ как инструменту познания человека в гносеологическом и онтологическом аспектах. Используя ИИ для своих целей, человек нагружает его данными (культурными пластами), оперируя которыми, ИИ выдаёт результат. Этот результат является интерпретацией ИИ исходных данных, то есть, тем, как он «видит» человека и его культуру. Исследуя эти результаты, человек может получить новые инструменты познания и лучше понять себя, увидев себя через оптику ИИ. С этим связана новая постановка вопросов о природе сознания – на стыке философии сознания, нейробиологии и кибернетики.

По итогам модуля планируется реализация студенческих проектов, которые предполагают работу с системами ИИ с целью создать творческие продукты и интерпретировать их с точки зрения разных культурологических и философских подходов. Будут проанализированы работы и главные идеи в интересующих нас областях ведущих современных философов и ученых, в том числе Стюарта Рассела и Питера Норвига, Дэвида Чалмерса, Ника Бострома, Эдварда Кастранова, Роджера Пенроуза, Ллойда Сета, Дэниела Деннета, Дэвида Дойча, Вадима Васильева и других.

2 модуль. ИИ как объект гуманитарного исследования. Преподаватель: В.Э. Терехович.

Люди рассуждают связными текстами на конкретном языке, используя специальные знаки. Процесс создания, чтения и интерпретации текстов – особая деятельность, она сложно устроена, упорядочена и подчинена правилам, а значит, в определенной степени рациональна. Генеративный ИИ (GenAI), основанный на Больших языковых моделях (LLM) тоже работает с текстами и ведет с их помощью диалог с человеком. Несмотря на то, что принципы работы GenAI задаются алгоритмами, сами тексты создаются им далеко не по формальному алгоритму. Это сложный пошаговый процесс, опирающийся как на контекст диалога, так и на огромный массив текстов, созданных другими людьми. Оказалось, что модели GenAI потенциально способны воспроизводить не только последовательности знаков, но и рассуждения, идеи и интерпретации, уже содержащиеся в человеческих текстах, к которым у GenAI есть доступ. Более того, GenAI быстро учиться моделям рассуждений, как с помощью людей, так и самостоятельно. Похоже ли это на то, как рассуждают люди в повседневных ситуациях, часто неосознанно воспроизводя образцы, ранее услышанного или прочитанного? Но тогда в какой степени имитация мышления человеком отличается от имитации GenAI? Чтобы ответить на эти вопросы, надо попробовать объяснить ход рассуждений GenAI, именно этому посвящены научные дискуссии об Explainable Artificial Intelligence (XAI).

Гуманитарии работают с текстами, созданными интеллектом человека. Поэтому в этом модуле мы попробуем взглянуть на генеративный ИИ одновременно и как на инструмент, и как на объект, благодаря изучению которого можно лучше понять механизм человеческого мышления. Генеративные модели ИИ появились недавно, поэтому мы сосредоточимся на обзоре самой современной дискуссии и классификации философских проблем, связанных с GenAI.

Для начала мы сами протестируем популярные модели GenAI и критически проверим популярные мифы о них. На основе научных публикаций попробуем разобраться в вопросе о том, существуют ли теоретические ограничения для способностей GenAI оперировать абстрактными понятиями и структурами, создавать новые идеи, рефлексировать, формулировать вопросы, интерпретировать контекст и скрытые смыслы. Обсудим принципиальные препятствия к проявлению ИИ агентности и субъектности, а также протестируем разные гипотезы в диалоге с ИИ.

Отдельно мы поговорим о другой важной теме: если ИИ способен эффективно имитировать отдельные элементы человеческого мышления, может ли он облегчить процесс научного исследования? Существуют ли непреодолимые препятствия для того, чтобы ИИ помогал выдвигать и сравнивать гипотезы, предсказывать явления для простых начальных условий, а также объяснять эти предсказания через причинные связи, теории и законы?

На основе материалов модуля студентам будет предложено участие в нескольких проектах.

3 модуль. Трансгуманизм и постгуманизм. Взаимодействие человеческих и нечеловеческих агентов. Преподаватель: Е.Н. Князева.

Рассмотрим дискуссии вокруг направлений гуманизма, трансгуманизма и постгуманизма в контексте идей мыслителей прошлого (Платон, Аристотель, Декарт, Лейбниц, Кант, Руссо, Ницше, Делёз), истории кибернетики, современных трендов в медицине и фармакологии, современных биомедицинских, информационных, когнитивных технологий по усилению работы человеческого тела и сознания (human enhancement). Изучим встречные движения: технологические изменения и модификации человеческого тела и сознания, способы построения технических устройств и роботов, приближенных к поведению человека, и способы взаимодействия humans и non-humans.

Современное движение трансгуманизма в его борьбе со старением и желаемым некоторыми достижением бессмертия (Р. Эттинджер) рассматривается в связи с идеями русских космистов (Н.Ф. Федоров, А.В. Сухово-Кобылин и др.). Радикальные трансгуманисты (Ник Бостром и Рэй Курцвейл) полагают, что значительные улучшения интеллекта потребуют от человека отказаться от биологии и передать мышление небиологическим платформам, например вычислительным устройствам. Современные возможности технологического изменения и конструирования человека включают в себя нейропротезирование, нейрокомпьютерный интерфейс, ноотропы, экзоскелет и т.п. Они быстро расширяются с перспективой появления гибридных или переходных существ, киборгов или пост-людей. Направление постгуманизма является более философски рефлексивным и включает в себя как технологический, так и нетехнологический аспекты. Термин «постчеловечество» (posthumanity) описывает либо новое сообщество разумных существ в виде суперчеловеческих, синтетических или гибридных (киборгов) организмов, которые появляются в результате технологической постгуманизации, либо саму социотехнологическую реальность нового типа, внутри которой происходит трансформация человеческой социальности (К. Хэйлз). Постгуманитарные исследования включают в себя множество направлений человеческого, не-человеческого и постчеловеческого миров: инвайронментализм (environmental humanities), анималистика (human animal studies), культурология новых медиа (cultural studies and new media), цифровые и технологические

гуманитарные исследования (digital and techno-humanities), медицинская гуманитаристика (medical humanities) и т.д.

Обсудим проблемы, связанные с построением ИИ, сложных технологических устройств, созданных на технологиях машинного обучения, роботов. Парадокс Моравека говорит о том, что устройства ИИ легче обучить выполнению сложных интеллектуальных задач, но сложно обучить процессам восприятия и мобильности, чем быстро овладевают годовалые человеческие младенцы. Обсуждается применение принципов отелесненного и энактивного познания (embodied and enacted cognition) в робототехнике и то, насколько технологические устройства и роботы могут быть наделены эмоциональностью и эмпатией. Стирание границ между естественным и искусственным ведет к опасности технологической сингулярности.

Студентам будет предложен проект по исследованию использования ИИ и больших данных в форсайте и построении сценариев развертывания событий на финансовом и экономических рынках.

4 модуль. Искусственный интеллект, наука больших данных и «новый эмпиризм». Преподаватель: Д.Н. Дроздова.

В последнем модуле мы остановимся детальнее на некоторых темах, которые уже затрагивались в предыдущих частях курса – а именно, на применении искусственного интеллекта в естественных, социальных и гуманитарных науках в работе с большими массивами данных. Некоторые авторы указывают, что последние десятилетия в науке произошел заметный сдвиг, вызванный появлением большого количества сырых данных, которые получают в реальном времени при помощи большого количества инструментов. Эти процессы позволяют говорить о появлении новой «парадигмы» в науке. При этом участие искусственного интеллекта в обработке этих данных и извлечении из них закономерностей является насущной потребностью.

Компьютерная обработка больших данных используется сейчас в физике и астрономии, медицине, экономике, социологии, метеорологии и многих других дисциплинах. С точки зрения эпистемологии, критерием правдоподобия результатов, получаемых при помощи технологий искусственного интеллекта, являются верифицированные предсказания. Но может ли искусственный интеллект не только предсказывать, но и формулировать теоретические гипотезы на основе больших массивов данных?

Мы особо остановимся в этом контексте на концепциях известного специалиста в области компьютерных технологий Джуды Перла, который настаивает, что только реабилитация в науке мышления, основанного на выявлении причинно-следственных связей, может способствовать увеличению теоретического содержания знания в области науки больших данных.

В конце мы также затронем проблемы этического характера, которые возникают в связи с ростом практики применения систем искусственного интеллекта в социально-экономических науках и медицине: в частности, вопросы использования персональных данных и ответственности технологий за прогнозы и предсказания, затрагивающие жизнь и благополучие человека.