Поведение машины

Предисловие

Машины, работающие на искусственном интеллекте (ИИ), все чаще опосредуют наши социальные, культурные, экономические и политические взаимодействия. Понимание поведения ИИ-систем необходимо для того, чтобы мы могли контролировать их действия, пожинать плоды и минимизировать вред. Мы утверждаем, что это требует широкой программы научных исследований для изучения поведения машин, которая включает в себя, но выходит за рамки дисциплины информатики и требует понимания со всех наук. Здесь мы сначала обозначим ряд вопросов, имеющих основополагающее значение для этой формирующейся области. Затем мы исследуем технические, правовые и институциональные ограничения, стоящие перед исследованием поведения машин.

Введение

В своей эпохальной книге 1969 года "науки об искусственном" 1 Нобелевский лауреат Герберт Саймон писал: "естественная наука-это знание о природных объектах и явлениях. Мы спрашиваем, не может ли существовать также "искусственная" наука—знание об искусственных объектах и явлениях."В соответствии с видением Саймона мы описываем появление новой междисциплинарной области научного исследования. Эта новая область связана с научными исследованиями интеллектуальных машин, но не как инженерных артефактов, а как нового класса акторов с определенными поведенческими паттернами и экологией. Эта область пересекается, но отличается от информатики и робототехники. Он рассматривает поведение машины эмпирически. Это сродни тому, как этология и поведенческая экология изучают поведение животных путем интеграции физиологии и биохимии-внутренних свойств-с изучением экологии и эволюции-свойств, формируемых окружающей средой. Поведение животных и человека не может быть полностью понято без изучения контекстов, в которых происходит поведение. Поведение машины также не может быть полностью понято без комплексного изучения алгоритмов и социальных сред, в которых работают алгоритмы.

В настоящее время ученые, изучающие поведение этих виртуальных и воплощенных агентов искусственного интеллекта (ИИ), в основном являются теми же учеными, которые создали самих агентов (повсюду мы используем термин “агенты ИИ” в широком смысле для обозначения как сложных, так и простых алгоритмов, используемых для принятия решений). Поскольку эти ученые создают агентов для решения конкретных задач, они чаще всего фокусируются на обеспечении того, чтобы агенты выполняли свою предназначенную функцию (хотя эти соответствующие области намного шире, чем конкретные примеры, перечисленные здесь). Например, агенты ИИ должны соответствовать критерию точности классификации документов, распознавания лиц или обнаружения визуальных объектов. Автономные автомобили должны успешно перемещаться в различных погодных условиях. Игровые агенты должны побеждать различных противников Человека или машины. И агенты интеллектуального анализа данных должны узнать, на каких людей ориентироваться в рекламных кампаниях в социальных сетях.

Эти агенты ИИ имеют потенциал, чтобы увеличить благосостояние и благополучие человека в мириады способов. Действительно, это типичное видение их создателей. Но более широкое рассмотрение поведения агентов ИИ теперь имеет решающее значение. Агенты ИИ будут все больше проникать в наше общество и уже участвуют во всем: от кредитного скоринга до алгоритмической торговли, от местной полиции до решений об условно-досрочном освобождении, от вождения до онлайн-знакомств до беспилотной войны 2,3. Комментаторы и ученые из различных областей, включая, но не ограничиваясь, когнитивную системную инженерию, взаимодействие человека с компьютером, человеческие факторы, научные технологии и общество, а также инженерия безопасности, поднимают тревогу по поводу широких, непреднамеренных последствий агентов ИИ, которые могут проявлять поведение и производить последующие социальные эффекты-как положительные, так и отрицательные-непредвиденные их создателями 4-7 .

В сочетании с этим отсутствием предсказуемости в отношении последствий ИИ существует опасение потенциальной потери человеческого контроля над интеллектуальными машинами 4 и потенциального вреда, связанного с увеличением использования машин для задач, когда-то выполняемых непосредственно людьми 8. В то же время исследователи описывают преимущества, которые агенты ИИ могут предложить обществу, поддерживая и усиливая принятие решений человеком 9,10. Хотя обсуждение этих вопросов привело ко многим важным выводам во многих отдельных областях академического исследования 11, при этом некоторые из них освещают проблемы безопасности автономных систем 12 или другие изучают последствия справедливости, подотчетности и прозрачности 13, многие вопросы остаются.

В этой обзорной статье рассматриваются и исследуются новые междисциплинарные области машинного поведения: научное исследование поведения интеллектуальных машин. Здесь мы очерчиваем ключевые темы исследований,вопросы и знаковые исследования, которые иллюстрируют новую область. Мы начнем с изучения поведения машин и обязательно междисциплинарного характера этой науки. Затем мы предоставляем основу для концептуализации исследований поведения машин. В заключение мы призываем к научному изучению машинной и человеко-машинной экологии и обсудим некоторые технические, правовые и институциональные барьеры, с которыми сталкивается эта область.

Мотивация к изучению поведения машин

Существует три основных мотивации для новой научной дисциплины-поведения машин. Во-первых, различные алгоритмы пронизывают наше общество и играть все возрастающую роль в нашей повседневной деятельности. Во-вторых, из-за сложных свойств этих алгоритмов и среды, в которой они работают, некоторые из их атрибутов и поведения могут быть трудно или невозможно формализовать аналитически. В-третьих, из-за их вездесущности и сложности прогнозирование воздействия интеллектуальных алгоритмов на человечество-как положительных, так и отрицательных-представляет собой серьезную проблему.

Повсеместность алгоритмов

Нынешняя Распространенность разнообразных алгоритмов в обществе беспрецедентна 4 (см. 1). Алгоритмы ранжирования новостей и боты социальных сетей влияют на информацию, которую видят граждане 14-18. Алгоритмы кредитного скоринга определяют кредитные решения 19-22. Алгоритмы онлайн-ценообразования формируют стоимость продукции дифференцированно по потребителям 23-25. Алгоритмической торговли программное обеспечение осуществляет операции на финансовых рынках c высокой скоростью 26-29. Алгоритмы формируют диспетчерскую и пространственную модели местной полиции 30, а программы алгоритмического назначения наказания влияют на время отбывания наказания в пенитенциарной системе 6. Автономные автомобили пересекают наши города, 31 в то время как алгоритмы совместного использования езды изменяют шаблоны поездок обычных транспортных средств 32. Машины картографируют наши дома, отвечая на словесные команды 33 и выполняя обычные бытовые задачи 34. Алгоритмы формируют романтические матчи через онлайн-знакомства 35,36. Машины, вероятно, будут все больше заменять людей в воспитании наших молодых 37 и уходе за нашими старыми 38 . И автономные агенты все чаще влияют на коллективное поведение, от групповой координации до совместного использования 39. Кроме того, хотя перспектива разработки автономного оружия является весьма спорной, поскольку многие на местах высказывают свое несогласие с этим решением, если такое оружие в конечном итоге будет развернуто, то машины смогут определить, кто живет, а кто умирает в вооруженном конфликте.

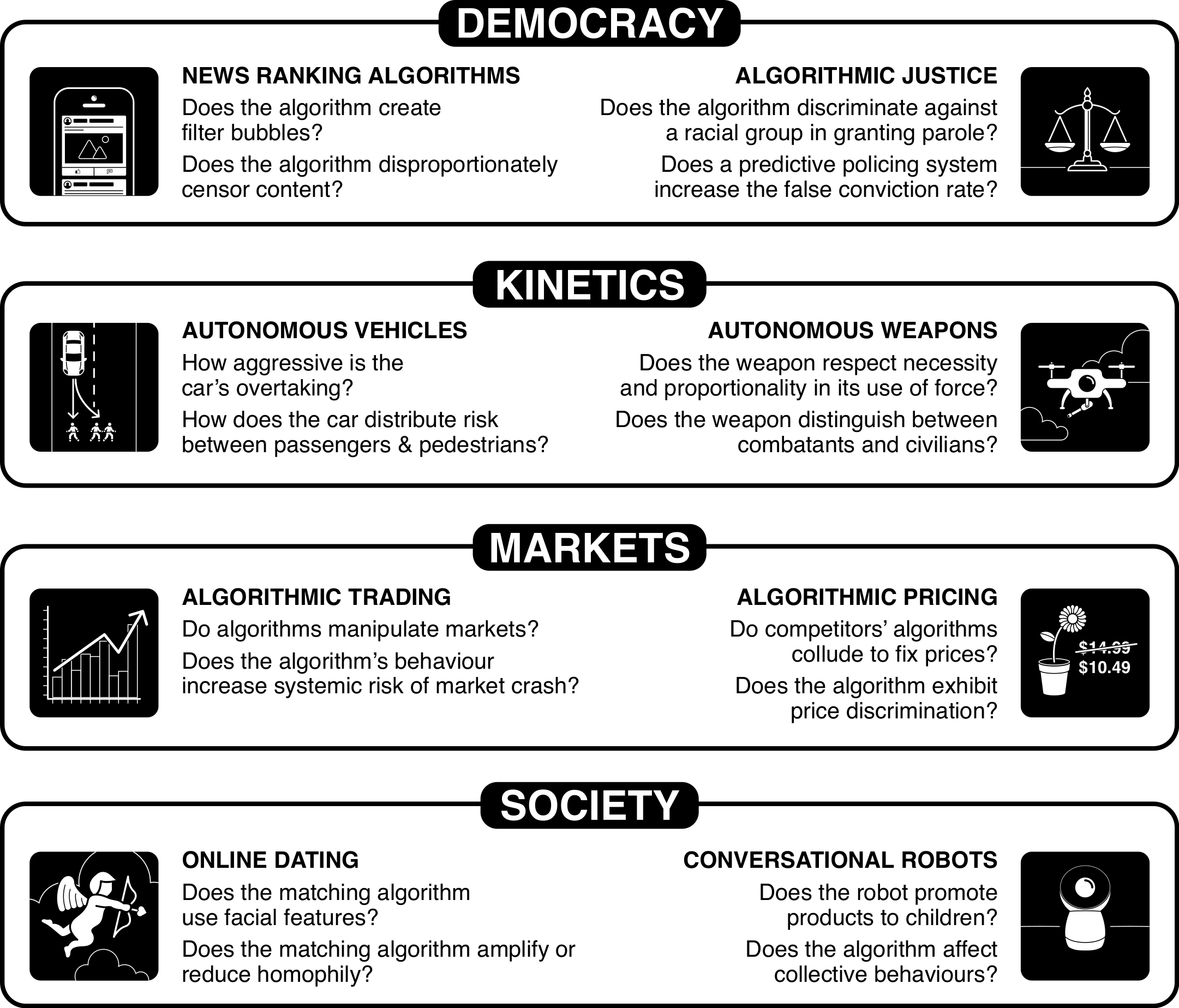


Рисунок 1: Примеры вопросов, которые относятся к области поведения машины.

Сложность и непрозрачность алгоритмов

Исключительное разнообразие этих ИИ-систем вкупе с их вездесущностью само по себе гарантировало бы, что изучение поведения таких систем будет представлять серьезную проблему, даже если сами отдельные алгоритмы будут относительно простыми. Сложность отдельных агентов ИИ в настоящее время высока и быстро возрастает. Хотя код для указания архитектур и обучения модели может быть простым, результаты могут быть очень сложными, часто эффективно приводящими к “черным ящикам” 43. Они поступают на вход и производят выход, но точный функциональный процесс генерации этого выхода трудно интерпретировать даже тем самым ученым, которые сами генерируют алгоритмы 44, хотя некоторый прогресс в интерпретируемости достигается 45,46. Кроме того, когда системы извлекают уроки из данных, их сбои связаны с несовершенством данных или методов сбора данных, что приводит к тому, что некоторые выступают за адаптацию механизмов отчетности для наборов данных 47 и моделей 48. Размерность и размер данных добавляют еще один уровень сложности к пониманию поведения машины 49.

Усугубляет эту проблему тот факт, что большая часть исходного кода и структуры моделей для наиболее часто используемых алгоритмов в обществе являются собственностью, как и данные, на которых эти системы обучаются. Промышленная тайна и правовая защита интеллектуальной собственности часто окружают исходный код и структуру модели. Во многих условиях единственными факторами, публично наблюдаемыми в промышленных системах ИИ, являются их входы и выходы.

Даже при наличии исходного кода или структуры модели агента ИИ может обеспечить недостаточную прогностическую власть над его выходом. Агенты ИИ могут также демонстрировать новое поведение через их взаимодействие с миром и другими агентами, которые невозможно предсказать с точностью 50. Даже если аналитические решения поддаются математическому описанию, они могут быть настолько длинными и сложными, что их невозможно расшифровать. Кроме того, когда окружающая среда меняется, возможно, в результате самого алгоритма, прогнозирование и анализ поведения становятся намного сложнее.

Благотворное и пагубное влияние алгоритмов на человечество

Повсеместное распространение алгоритмов в сочетании с их возрастающей сложностью, как правило, усиливает трудности оценки влияния алгоритмов на людей и общество. Агенты ИИ могут формировать поведение человека и социальные результаты как намеренно, так и непреднамеренно. Например, некоторые агенты ИИ предназначены для оказания помощи в обучении детей 53, а другие - для оказания помощи пожилым людям 38,54 . Эти системы ИИ могут принести пользу их предназначенным людям, подталкивая этих людей к лучшему обучению или более безопасному поведению мобильности. Тем не менее, с возможностью подтолкнуть человеческое поведение позитивным или намеренным образом возникает риск того, что человеческое поведение может быть подтолкнуто дорогостоящим или непреднамеренным образом-дети могут быть под влиянием купить определенные фирменные продукты, а старшие могут быть подтолкнуты к просмотру определенных телевизионных программ.

То, как такое алгоритмическое влияние на отдельных людей масштабируется в общесоциальные воздействия, как положительные, так и отрицательные, вызывает критическую озабоченность. Например, воздействие политической дезинформации на небольшое число людей может оказать незначительное воздействие на общество в целом. Однако эффект внедрения и распространения такой дезинформации в социальных сетях может иметь более существенные социальные последствия 55-57. Кроме того, вопросы алгоритмической справедливости или предвзятости 58,59 уже были задокументированы в различных контекстах, включая компьютерное зрение 60, встраивание слов 61,62, рекламу 63, полицию 64, уголовное правосудие 6,65 и социальные услуги 66. Для решения этих проблем практикующие иногда будут вынуждены идти на ценностный компромисс между конкурирующими и несовместимыми понятиями предвзятости или между предубеждениями человека и машины. Дополнительные вопросы, касающиеся влияния алгоритмов, включают: как алгоритмы онлайн-знакомств изменяют социальный институт брака 35,36? Существуют ли системные эффекты усиления взаимодействия с интеллектуальными алгоритмами на этапах и скорости развития человека 53? Эти вопросы усложняются в "гибридных системах", состоящих из множества машин и людей, взаимодействующих и проявляющих коллективное поведение. Для того чтобы общество могло вносить свой вклад в ИИ и контролировать его последствия, Ученые, изучающие поведение машин, должны дать представление о том, как работают эти системы, а также о выгодах, затратах и компромиссах, связанных с повсеместным использованием ИИ в обществе.

Междисциплинарное исследование поведения машин

Для изучения поведения машин-особенно поведения алгоритмов черного ящика в реальных условиях - мы должны интегрировать знания из множества научных дисциплин (см. 2). Эта интеграция в настоящее время находится на зарождающейся стадии и происходит в основном на разовой основе в ответ на растущую потребность в понимании поведения машин. В настоящее время ученые, которые чаще всего изучают поведение машин, - это в первую очередь компьютерщики, робототехники и инженеры, создающие эти машины. Эти ученые могут быть опытными математиками и инженерами, но, как правило, не являются обученными бихевиористами. Они редко получают формальные инструкции по экспериментальной методологии, демографической статистике и парадигмам выборки или наблюдательному причинно-следственному выводу, не говоря уже о нейробиологии, коллективном поведении или социальной теории. И наоборот, в то время как ученые-бихевиористы с большей вероятностью будут обучаться этим научным методам, они с меньшей вероятностью будут обладать знаниями, необходимыми для того, чтобы квалифицированно оценить качество и уместность методов ИИ для данной проблемной области или математически описать свойства конкретных алгоритмов.

Интегрировать научную практику из разных областей непросто. До этого момента основное внимание тех, кто создает системы ИИ по разработке, реализации и оптимизации интеллектуальных систем для выполнения специализированных задач. Отличный прогресс был достигнут в тестовых задачах, от настольных игр, таких как шахматы 68, шашки 69 и Go 70,71 до карточных игр, таких как покер 72, до компьютерных игр, таких как на платформе Atari 73, до искусственных рынков 74, Robocup Soccer 75, а также стандартизированных данных оценки, таких как данные ImageNet для распознавания объектов 76 и общие объекты Microsoft в контекстных данных для задач подписи изображений 77. Успех также был достигнут в распознавании речи, переводе языка и автономном перемещении. Эти критерии сочетаются с метриками для количественной оценки производительности по стандартизированным задачам 78-81 и используются для улучшения производительности, прокси-сервера, который позволяет разработчикам ИИ стремиться к лучшим, быстрым и более надежным алгоритмам.

Но методологии, направленные на максимизацию алгоритмической производительности, не являются оптимальными для проведения научных наблюдений за свойствами и поведением агентов ИИ. Вместо того чтобы использовать метрики для оптимизации по сравнению с эталонами, ученые, изучающие поведение машин, заинтересованы в более широком наборе показателей, подобно тому как социологи исследуют широкий спектр человеческого поведения в области социального, политического или экономического взаимодействия. Поэтому ученые, изучающие поведение машин, прилагают значительные усилия для определения новых показателей микро-и макро-результатов, позволяющих ответить на широкие вопросы. Как эти алгоритмы могут вести себя в различных средах? Как взаимодействие человека с этими алгоритмами может изменить социальные результаты? Рандомизированные эксперименты, выводы наблюдений и демографическая описательная статистика-методы, широко используемые в количественных науках о поведении-должны быть центральными для изучения поведения машин. Включение ученых извне дисциплин, которые традиционно производят интеллектуальные машины, может дать знания о важных методологических инструментах, научных подходах, альтернативных концептуальных рамках и новых перспективах экономических, социальных и политических явлений, на которые машины будут все больше влиять.

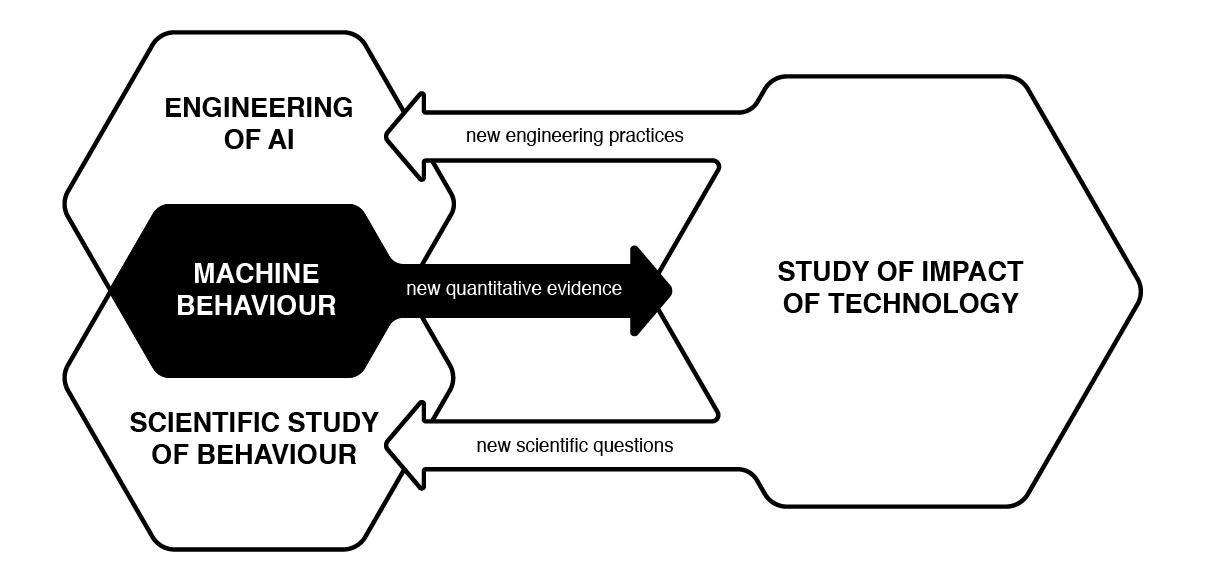


Рисунок 2: междисциплинарность поведения машины.

Поведение машины находится на пересечении областей, которые проектируют и проектируют системы ИИ, и областей, которые традиционно используют научный метод для изучения поведения биологических агентов. Результаты исследований поведения машин дают новые количественные данные, которые могут помочь информировать те области, которые изучают потенциальное воздействие технологии на социальные и технологические системы. В свою очередь, эти области могут дать полезную новую инженерную практику и новые научные вопросы областям, изучающим поведение машин. Наконец, научное изучение поведения помогает исследователям ИИ делать более точные заявления о том, что могут и чего не могут делать ИИ-системы.

Тип вопроса и объект исследования

Николаас Тинберген, получивший в 1973 году Нобелевскую премию по физиологии и медицине наряду с Карлом фон Фришем и Конрадом Лоренцем за создание области этологии, выделил четыре дополнительных измерения анализа для объяснения поведения животных. Эти измерения касаются вопросов функции, механизма, развития и эволюционной истории поведения и обеспечивают организационную основу для изучения поведения животных и человека. Например, эта концептуализация отличает изучение того, как молодое животное или человек развивает поведение, от эволюционной траектории, выбранной для такого поведения в популяции. Цель этих различий - не разделение, а интеграция. Хотя не будет ошибкой сказать, что, например, песня птицы объясняется обучением или ее конкретной эволюционной историей, полное понимание песни потребует и того, и другого.

Несмотря на фундаментальные различия между машинами и животными, изучение поведения машин может выиграть от подобной классификации. Машины имеют механизмы, которые производят поведение, проходят развитие, которое интегрирует экологическую информацию в поведение, производят функциональные последствия, которые заставляют конкретные машины становиться более или менее распространенными в конкретных средах, и воплощают эволюционную историю, через которую прошлые среды и человеческие решения продолжают влиять на поведение машин. Ученые в области информатики уже достигли значительных успехов в понимании механизмов и развития систем ИИ,хотя многие вопросы остаются. Относительно меньше внимания уделяется функциям и эволюции систем ИИ. Мы обсудим эти четыре темы В следующих подразделах и представим диаграмму 3 в виде резюме 84.

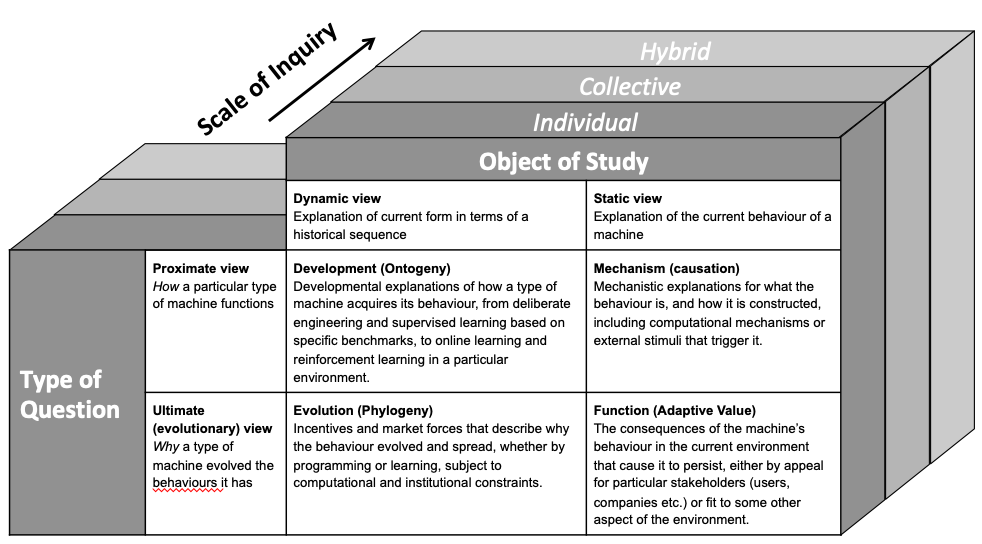


Рисунок 3: тип вопроса и Объект исследования Тинбергена, модифицированный для изучения поведения машины. Четыре категории, предложенные Тинбергеном для изучения поведения животных, могут быть адаптированы к изучению поведения машин 83,84. Структура Тинбергена предлагает два типа вопросов: как и почему, а также два взгляда на эти вопросы: динамический и статический. Каждый вопрос может быть рассмотрен на трех уровнях исследования: отдельные машины, коллективы машин и гибридные человеко-машинные системы.

Механизм формирования поведения

Непосредственные причины поведения машины связаны с тем, как поведение наблюдательно запускается и генерируется в определенных средах. Например, ранние алгоритмические торговые программы использовали простые правила для запуска поведения покупки и продажи 85. Более сложные агенты могут вычислять стратегии, основанные на адаптивной эвристике или явной максимизации ожидаемой полезности 86. Поведение алгоритма обучения подкреплению, который играет в покер, можно отнести к тому, как он представляет пространство состояний или оценивает игровое дерево 72 и т. д.

Механизм зависит как от алгоритма, так и от его среды. Более сложный агент, такой как автомобиль без водителя, может демонстрировать особое поведение за рулем-например, переключение полосы движения, обгон, сигнализация пешеходам. Это поведение будет генерироваться в соответствии с алгоритмами, которые строят политику вождения 87, но также формируются в основном особенностями системы восприятия и приведения в действие автомобиля, включая разрешение и точность его системы обнаружения и классификации объектов и отзывчивость и точность его рулевого управления, среди других факторов. Поскольку многие из современных систем ИИ являются производными от методов машинного обучения (мл), применяемых ко все более сложным данным, изучение механизма поведения машины, такого как упомянутые выше, потребует продолжения работы над методами интерпретируемости для мл 46,88,89.

Развитие поведения

При изучении поведения животных или человека под развитием понимается то, как индивид приобретает определенное поведение, например, посредством имитации или обусловленности окружающей средой. Это отличается от долгосрочных эволюционных изменений.

В контексте машин мы можем спросить, как машины приобретают (развивают) конкретное индивидуальное или коллективное поведение. Развитие поведения может быть напрямую связано с выбором человека-инженерным или проектным. Выбор архитектурного дизайна, сделанный программистом (например, значение параметра скорости обучения, получение представления знаний и состояния или конкретная проводка сверточной нейронной сети), определяет или влияет на виды поведения, которые демонстрирует алгоритм. В более сложной системе ИИ, такой как автомобиль без водителя, поведение автомобиля развивается со временем из разработки программного обеспечения и изменения аппаратных компонентов, которые инженеры включают в свою общую архитектуру. Поведение также может измениться в результате алгоритмических обновлений, подталкиваемых к машине ее разработчиками после развертывания.

Инженер-человек может также формировать поведение машины, подвергая ее определенным тренировочным стимулам. Например, многие алгоритмы классификации изображений и текстов обучены оптимизации точности для определенного набора наборов данных с человеческими метками. Выбор набора данных - и тех функций, которые он представляет 60,61 - может существенно повлиять на поведение, отображаемое алгоритмом.

Наконец, машина может приобрести поведение через собственный опыт. Например, агент по обучению подкреплению, обученный максимизировать долгосрочную прибыль, может изучить специфические краткосрочные торговые стратегии, основанные на его собственных прошлых действиях и сопутствующей обратной связи с рынком . Аналогичным образом, алгоритмы рекомендаций продуктов делают рекомендации, основанные на бесконечном потоке решений, сделанных клиентами, соответствующим образом обновляя свои рекомендации.

Функция

В области поведения животных адаптивная ценность описывает, как поведение способствует пожизненной репродуктивной пригодности животного. Например, определенное охотничье поведение может быть более или менее успешным, чем другое, в продлении жизни животного и, соответственно, числа возможностей спаривания, в результате чего рождается потомство и вероятный репродуктивный успех потомства. Акцент на функции помогает понять, почему одни поведенческие механизмы распространяются и сохраняются, в то время как другие снижаются и исчезают. Функция зависит критически на пригонке к окружающей среде.

В случае с машинами, мы можем говорить о том, как поведение выполняет одновременно функцию для конкретных заинтересованных сторон человека. Человеческая среда создает селективные силы, которые могут сделать некоторые машины более распространенными. Успешные модели поведения (улучшающие” приспособленность") копируются разработчиками другого программного и аппаратного обеспечения или иногда проектируются для распространения среди самих машин. Эта динамика в конечном счете определяется успехом учреждений-корпораций, больниц, муниципальных органов управления, университетов и т. д. -- что построить или использовать ИИ. Наиболее очевидным примером является алгоритмическая торговля, в которой успешные автоматические торговые стратегии могут быть скопированы по мере того, как их разработчики переходят от компании к компании или просто наблюдаются и реверсируются конкурентами.

Эти силы могут производить непредвиденные эффекты. Например, такие цели, как максимизация вовлеченности на сайте социальных сетей, могут привести к так называемым "пузырям фильтров" 91, которые могут усилить политическую поляризацию или без тщательной умеренности могут способствовать распространению поддельных новостей. Тем не менее, веб-сайты, которые не оптимизируются для взаимодействия с пользователями, могут быть не столь успешными по сравнению с теми, которые это делают, или могут вообще выйти из бизнеса. Аналогичным образом, в отсутствие внешнего регулирования автономные автомобили, которые не уделяют приоритетного внимания безопасности своих пассажиров, могут быть менее привлекательными для потребителей, что приводит к сокращению продаж 31. Иногда функция поведения машины состоит в том, чтобы справляться с поведением других машин. Враждебные атаки-синтетические входы, которые обманывают систему, создавая нежелательный выход 44,92-94 - на системы ИИ и последующие реакции тех, кто развивает ИИ на эти атаки 95, могут привести к сложной динамике хищника-жертвы, которую нелегко понять, изучая каждую машину в отдельности.

Эти примеры показывают, как стимулы, создаваемые внешними институтами и экономическими силами, могут оказывать косвенное, но значительное влияние на поведение машин . Понимание взаимодействия между этими стимулами и ИИ имеет отношение к изучению поведения машины. Эта рыночная динамика, в свою очередь, будет взаимодействовать с другими процессами, чтобы произвести эволюцию среди машин и алгоритмов.

Эволюция

В исследовании поведения животных филогения описывает, как развивалось поведение. В дополнение к текущей функции на поведение влияют прошлые избирательные давления и ранее разработанные механизмы. Например, человеческая рука возникла из плавника костистой рыбы. Его текущая функция больше не для плавания, но его внутренняя структура объясняется его эволюционной историей. Неизбирательные силы, такие, как миграция и дрейф, также играют важную роль в объяснении взаимосвязей между различными формами поведения.

В случае машин эволюционная история также может порождать зависимость от пути, объясняя в противном случае загадочное поведение. На каждом этапе аспекты алгоритмов повторно используются в новых контекстах, сдерживая будущее поведение и делая возможными дополнительные инновации. Например, ранний выбор микропроцессорного дизайна продолжает влиять на современные вычисления, а традиции проектирования алгоритмов—нейронные сети и байесовские модели пространства состояний, например,-строят многие предположения и направляют будущие инновации, делая некоторые новые алгоритмы более легкими для доступа, чем другие. В результате некоторые алгоритмы могут обращать внимание на определенные функции и игнорировать другие, поскольку эти функции были важны в ранних успешных приложениях. Поведение некоторых машин может распространяться, потому что оно” эволюционирует", легко модифицируется и устойчиво к возмущениям, подобно тому, как некоторые черты животных могут быть общими, поскольку они способствуют разнообразию и стабильности .

Конечно, поведение машины развивается совершенно иначе, чем поведение животных. Большинство животных наследуют просто-два родителя, одно событие передачи. Алгоритмы гораздо более гибкие, и у них есть дизайнер с целью в фоновом режиме. Человеческая среда сильно влияет на эволюцию алгоритмов, изменяя их систему наследования. Поведение репликации ИИ может быть облегчено с помощью культуры совместного использования программного обеспечения с открытым исходным кодом, деталей сетевой архитектуры или базовых учебных наборов данных. Например, компании, которые разрабатывают программное обеспечение для автомобилей без водителя, могут совместно использовать расширенные библиотеки с открытым исходным кодом для обнаружения объектов или планирования пути, а также учебные данные, лежащие в основе этих алгоритмов, чтобы обеспечить распространение программного обеспечения для повышения безопасности по всей отрасли. Одна адаптивная "мутация" в поведении конкретного автомобиля без водителя может мгновенно распространяться на миллионы других автомобилей посредством обновления программного обеспечения. Однако другие учреждения также применяют ограничения. Например, патенты на программное обеспечение могут накладывать ограничения на копирование определенных поведенческих черт. И нормативные ограничения - такие как законы О защите конфиденциальности - могут помешать машинам получать доступ, сохранять или иным образом использовать определенную информацию в процессе принятия решений. Эти особенности подчеркивают, что машины могут демонстрировать очень разные эволюционные траектории, поскольку они не связаны механизмами органической эволюции.

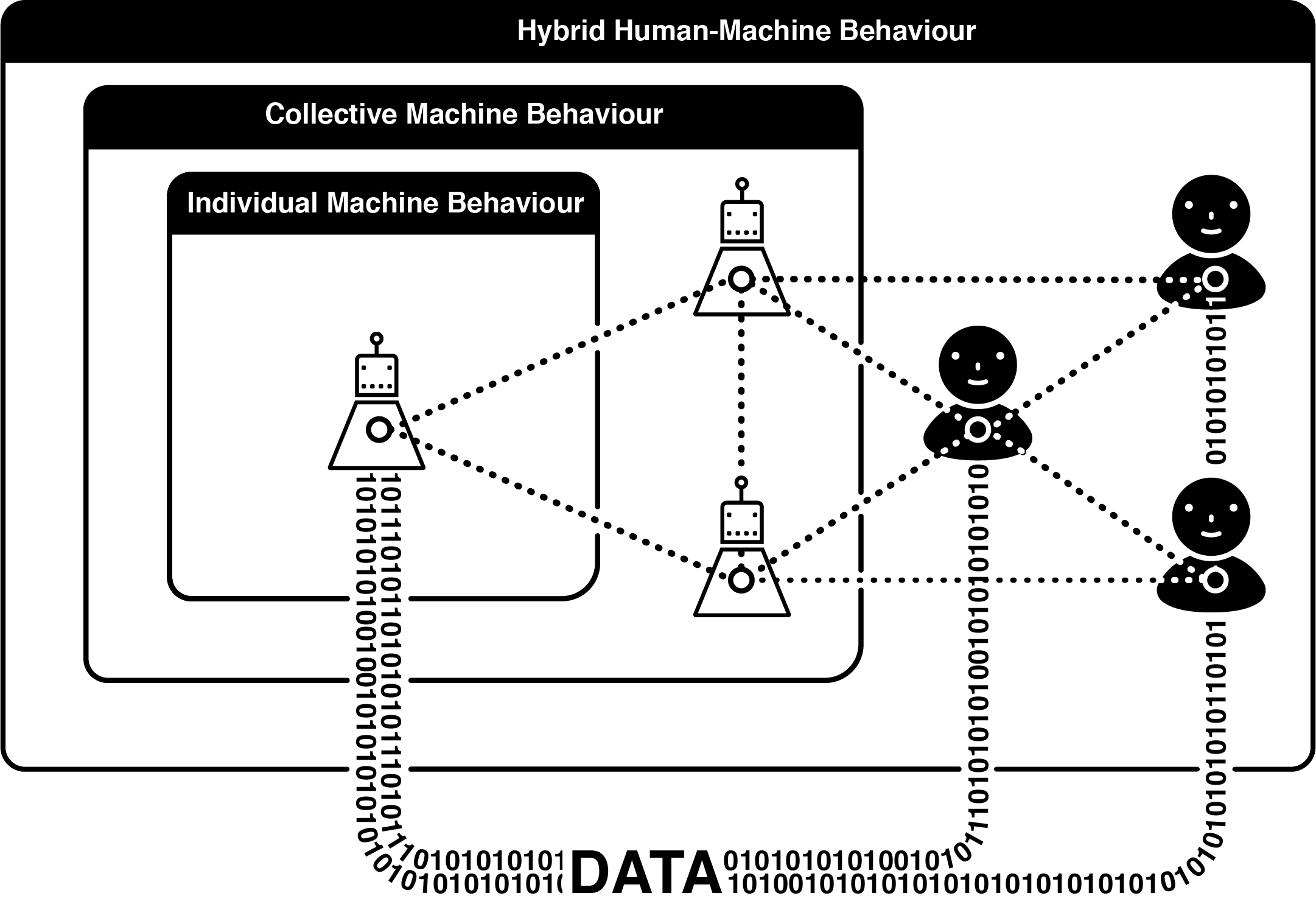


Рисунок 4: масштаб исследования в экосистеме поведения машины. Системы ИИ представляют собой объединение людей, данных и алгоритмов. Каждая из этих областей влияет на другую как хорошо понятными, так и еще неизвестными способами. Данные, отфильтрованные с помощью алгоритмов, созданных людьми, влияют на индивидуальное и коллективное поведение машины. Системы ИИ обучаются данным, в свою очередь влияя на то, как люди генерируют новые данные. Системы ИИ коллективно взаимодействуют друг с другом и влияют друг на друга. Человеческие взаимодействия могут быть изменены введением этих систем ИИ. Исследования поведения машин, как правило, происходят на индивидуальной, коллективной или гибридной человеко-машинной шкале исследований.

Шкала расследования

В рамках, описанных выше и на рис. 3, мы теперь каталогизируем примеры поведения машин на трех шкалах исследования: отдельные машины, коллективы машин и группы машин, встроенных в социальную среду с группами людей в “гибридных” или “гетерогенных” системах (см. 4).Особое внимание уделяется изучению самого алгоритма, коллективное поведение машин - изучению взаимодействий между машинами, а гибридное поведение человека и машины-изучению взаимодействий между машинами и людьми. Здесь мы можем провести аналогию с изучением конкретного вида, изучением взаимодействий между членами вида и взаимодействий вида с их более широкой средой обитания. Анализ на любой из этих шкал может касаться любого или всех вопросов, указанных на диаграмме 3.

Индивидуальное поведение машины

Изучение индивидуального поведения машины фокусируется на конкретных интеллектуальных машинах сами по себе. Часто эти исследования фокусируются на свойствах, присущих отдельным машинам, свойствах, обусловленных их исходным кодом или дизайном. В настоящее время большинство этих исследований проводится в области машинного обучения и программной инженерии. Существует два общих подхода к изучению поведения отдельных машин. Первый фокусируется на профилировании набора поведений любого конкретного машинного агента с использованием внутримашинного подхода, сравнивая поведение конкретной машины в разных условиях. Второй подход, межмашинный, исследует, как различные отдельные агенты машины ведут себя в одном и том же состоянии.

Внутримашинный подход к изучению поведения отдельных машин задает такие вопросы, как: существуют ли константы, характеризующие внутримашинное поведение любого конкретного ИИ в различных контекстах? Как поведение конкретного ИИ развивается с течением времени в той же или иной среде? Какие факторы окружающей среды приводят к выражению определенного поведения машинами?

Например, алгоритм может только демонстрировать определенное поведение, если обучение на конкретных исходных данных 98-100 (разработка, Рис. 3). Ведет ли себя алгоритм, оценивающий вероятность рецидива в решениях об условно-досрочном освобождении 6, неожиданным образом при представлении оценочных данных, существенно отличающихся от его обучающих данных? Другие исследования, связанные с характеристикой поведения внутри машины, включают изучение индивидуального поведения роботов-роботов 101,102, "когнитивных" атрибутов алгоритмов и полезности использования методов психологии при изучении алгоритмического поведения 103, а также изучение специфических характеристик ботов, таких как те, которые предназначены для "влияния" на пользователей человека 104.

Второй подход к изучению поведения отдельных машин рассматривает то же поведение, что и у разных машин. Например, те, кто заинтересован в изучении рекламного поведения интеллектуальных агентов, могут исследовать различные рекламные платформы (и лежащие в их основе алгоритмы) и исследовать межмашинный эффект от проведения экспериментов с одним и тем же набором рекламных входов на разных платформах. Такой же подход может быть принят для исследований алгоритмов динамического ценообразования 23,24,32 на разных платформах. В других межмашинных исследованиях можно было бы рассмотреть различные модели поведения, используемые автономными транспортными средствами при обгоне, или различные модели поведения при поиске и спасении, демонстрируемые беспилотными летательными аппаратами 107.

Коллективное поведение машины

В отличие от индивидуального поведения машины, коллективное поведение машины фокусируется на интерактивном и общесистемном поведении коллекций агентов машины. В некоторых случаях последствия поведения отдельных машин могут иметь мало смысла до тех пор, пока не будет рассмотрен коллективный уровень. Некоторые исследования этих систем были вдохновлены естественными коллективами, такими как рои насекомых, или мобильными группами, такими как стаи птиц или стайки рыб. Например, известно, что группы животных демонстрируют как эмерджентное восприятие сложных экологических особенностей 108, так и эффективное принятие консенсусных решений 109 . В обоих сценариях группы демонстрируют осведомленность об окружающей среде, которая не существует на индивидуальном уровне. Такие области, как мультиагентные системы и теория вычислительных игр, дают полезные примеры изучения этой области поведения машин.

Роботы, которые используют простые алгоритмы для локальных взаимодействий между ботами, тем не менее могут производить интересное поведение после объединения в большие коллективы. Например, ученые исследовали Роя свойств микро-роботов, которые объединяются в скопления, напоминающие стаи в системах биологических агентов 110,111. Дополнительные примеры включают коллективное поведение алгоритмов как в лаборатории (в игре жизни 112), так и в дикой природе (как видно из Википедии редактирования ботов 113). К числу других примеров можно отнести появление новых алгоритмических языков между взаимодействующими интеллектуальными машинами, а также динамические свойства полностью автономных транспортных систем. В конечном счете, многие интересные вопросы в этой области еще предстоит изучить.

Подавляющее большинство работ по коллективному поведению животных и коллективной робототехнике сосредоточено на том, как взаимодействия между простыми агентами могут создавать структуры и свойства более высокого порядка. Хотя это и важно, это игнорирует тот факт, что многие организмы, а также агенты ИИ 75, являются сложными сущностями, чье поведение и взаимодействия могут не характеризоваться упрощенными представлениями. Выявление того, какие дополнительные свойства возникают, когда взаимодействующие сущности способны к сложному познанию, остается ключевой задачей биологических наук и может иметь прямые параллели в изучении поведения машин. Например, подобно животным, машины могут демонстрировать "социальное обучение". Такое социальное обучение не должно ограничиваться обучением машин у машин, но мы можем ожидать, что машины будут учиться у людей, и наоборот, люди будут учиться на поведении машин. Внедренные процессы обратной связи могут коренным образом изменить накопление знаний, в том числе на протяжении поколений, оказывая непосредственное воздействие на человеческую и машинную “культуру”.

Кроме того, созданные человеком системы ИИ не обязательно сталкиваются с теми же ограничениями, что и организмы, а коллективные сборки машин обеспечивают новые возможности, такие как мгновенная глобальная коммуникация, которая может привести к совершенно новым коллективным поведенческим моделям. Исследования коллективного поведения машин исследуют свойства сборок машин, а также неожиданные свойства, которые могут возникнуть из этих сложных систем взаимодействий.

Например, некоторые из наиболее интересных коллективных алгоритмов поведения наблюдались в финансовых торговых средах. Эти среды работают в крошечных временных масштабах, так что алгоритмические трейдеры могут реагировать на события и друг на друга раньше любого человеческого трейдера. При определенных условиях высокочастотные возможности могут привести к неэффективности на финансовых рынках 26 115. В дополнение к беспрецедентной скорости реагирования широкое использование машинного обучения, автономной работы и способности развертываться в масштабе-все это основания полагать, что коллективное поведение машинной торговли может качественно отличаться от поведения людей. Кроме того, эти финансовые алгоритмы и торговые системы обязательно обучаются определенным наборам исторических данных и реагируют на ограниченное разнообразие прогнозируемых сценариев. Как они будут реагировать на ситуации, которые являются новыми и непредвиденными в их конструкции? Flash-сбои являются примерами явно непреднамеренных последствий (взаимодействующих) алгоритмов 116,117. Могут ли алгоритмы взаимодействовать для создания более крупного рыночного кризиса?

Гибридное поведение человека и машины

Люди все чаще взаимодействуют с машинами 16. Они опосредуют наши социальные взаимодействия, формируют новости, которые мы видим, и формируют отношения с нами, которые могут изменить наши социальные системы. Из-за своей сложности эти гибридные человеко-машинные системы представляют собой одну из наиболее технически сложных и одновременно наиболее важных областей изучения поведения машин.

Машины формируют поведение человека

Одна из наиболее очевидных, но тем не менее жизненно важных областей исследования поведения машин связана с тем, как внедрение интеллектуальных машин в социальные системы может изменить убеждения и поведение людей. Как и при внедрении автоматизации в производственный процесс 119, интеллектуальные машины могут создавать новые социальные проблемы в процессах совершенствования существующих проблем. Изменяют ли алгоритмы сопоставления, используемые для онлайн-знакомств, распределительные результаты процесса знакомств? Изменяют ли алгоритмы фильтрации новостей распределение общественного мнения? Могут ли небольшие ошибки в алгоритмах или данные, которые они используют для создания воздействия на общество? Как интеллектуальные роботы в наших школах, больницах и центрах по уходу изменяют развитие человека и качество жизни и влияют на результаты для инвалидов?

Другие вопросы в этой области связаны с потенциалом машин для изменения социальной структуры более фундаментальными способами. В какой степени и каким образом правительства используют машинный интеллект для изменения характера демократии, политической подотчетности и транспарентности или гражданского участия? В какой степени интеллектуальные машины влияют на деятельность полиции, наблюдение и ведение войны? Насколько большое влияние оказали боты на итоги выборов 56? Могут ли системы ИИ, помогающие в формировании человеческих социальных отношений, способствовать коллективным действиям?

Важно отметить, что исследования в этой области также изучают, как люди воспринимают использование машин как средство принятия решений 7123, предпочтения людей за и против использования алгоритмов 124 и степень, в которой человекоподобные машины производят или уменьшают дискомфорт у людей 39 125. Важным вопросом в этой области является то, как люди реагируют на растущее совместное производство экономических товаров и услуг в тандеме с интеллектуальными машинами. В конечном счете, понимание того, как человеческие системы могут быть изменены введением интеллектуальных машин в нашу жизнь, является жизненно важным компонентом изучения поведения машин.

Люди формируют поведение машины

В то время как разумные машины могут изменять человеческое поведение, люди также создают, информируют и формируют поведение разумных машин. Мы формируем машинное поведение посредством непосредственного проектирования ИИ-систем и посредством обучения этих систем как активному человеческому входу, так и пассивному наблюдению человеческого поведения с помощью данных, которые мы создаем ежедневно. Выбор того, какие алгоритмы использовать, какую обратную связь предоставлять этим алгоритмам и на основе каких данных их обучать, также в настоящее время является человеческим решением и может непосредственно изменять поведение машины. Важным компонентом в изучении поведения машины является понимание того, как эти технологии изменяют результирующее поведение ИИ. Отвечают ли обучающие данные за конкретное поведение машины? Это сам алгоритм? Или это какая-то комбинация алгоритма и данных? Рамки, изложенные на рис. 3, предполагают наличие дополнительных ответов на каждый из этих вопросов. Изучение того, как изменение параметров инженерного процесса может изменить последующее поведение интеллектуальных машин при взаимодействии с другими машинами и людьми в дикой природе, имеет центральное значение для целостного понимания поведения машин.

"Человек-машина" совместное поведение

Хотя методологически удобно отделять исследования о том, как люди формируют машины и наоборот, большинство систем ИИ функционируют в областях, где они сосуществуют с людьми в сложных гибридных системах 67,39,125,128. Вопросы, имеющие важное значение для изучения этих систем, включают те, которые изучают поведение, характеризующее взаимодействие человека и машины, включая сотрудничество, конкуренцию и координацию. Например, как человеческие предубеждения в сочетании с ИИ могут изменить человеческие эмоции или убеждения 14,55,56,129,130 ? Как человеческие тенденции могут сочетаться с алгоритмами, способствующими распространению информации? Как можно изменить структуру движения на улицах, населенных большим количеством автомобилей без водителя и автомобилей, управляемых людьми? Как могут быть изменены торговые шаблоны взаимодействиями между людьми и алгоритмическими торговыми агентами? И какие факторы могут способствовать доверию и сотрудничеству между людьми и машинами?

Другая тема в этой области касается роботизированной и программно-управляемой автоматизации человеческого труда 132 . Здесь мы видим два различных типа взаимодействия машины и человека. Одна из них заключается в том, что машины могут повысить эффективность человека, например, в роботизированной и компьютерной хирургии. Другое что машины могут заменить людей, в перевозке без водителя и доставке пакета. Будут ли машины в конечном итоге делать больше замены или улучшения в долгосрочной перспективе? Что человека-машину Co-поведение приведет?

Приведенные выше примеры свидетельствуют о том, что во многих вопросах, касающихся гибридного поведения человека и машины, необходимо рассматривать циклы обратной связи между влиянием человека на поведение машины и влиянием машины на поведение человека одновременно. Ученые начали изучать взаимодействие человека и машины в формальной лабораторной среде, наблюдая, что взаимодействие с простыми ботами может повысить человеческую координацию и что боты могут напрямую сотрудничать с людьми на уровнях, конкурирующих с человеческим сотрудничеством 133. Однако по-прежнему существует острая необходимость в дальнейшем понимании циклов обратной связи в дикой природе, где люди все чаще используют алгоритмы для принятия решений 134 и впоследствии информируют обучение тем же алгоритмам через эти решения. Кроме того, по всем типам вопросов в области машинной поведенческой экологии существует необходимость в исследованиях, которые изучают долгосрочную динамику этих гибридных систем 53 с особым упором на то, как социальные взаимодействия человека 135 136 могут быть изменены путем внедрения интеллектуальных машин 137.

Прогноз

Дальнейшее изучение поведения машин имеет решающее значение для максимизации потенциальных выгод ИИ для общества. Последовательный выбор, который мы делаем в отношении интеграции агентов ИИ в человеческую жизнь, должен быть сделан с некоторым пониманием возможных социальных последствий этого выбора. Чтобы обеспечить это понимание и предвидение, нам нужна новая междисциплинарная область научных исследований: поведение машины.

Для успешной работы в этой области существует ряд важных соображений. Во-первых, изучение поведения машин не означает, что алгоритмы ИИ обязательно должны иметь независимый орган, и не означает, что алгоритмы должны нести моральную ответственность за свои действия. Если собака кусает прохожего, ответственность несет владелец собаки. Тем не менее, полезно изучить поведенческие модели животных, чтобы предсказать такое аберрантное поведение. Машины работают в рамках более широкой социально-технической структуры, и их человеческие участники в конечном счете несут ответственность за любой вред, который может причинить их развертывание.

Во-вторых, некоторые комментаторы могут предположить, что рассмотрение систем ИИ в качестве агентов закрывает фокус на базовых данных, на которых обучаются такие системы ИИ. Действительно, ни одно поведение никогда полностью не отделимо от экологических данных, на основе которых этот агент обучается или развивается; поведение машины не является исключением. Однако столь же важно понять, как поведение машин изменяется в зависимости от измененных факторов окружающей среды, как и то, как поведение биологических агентов меняется в зависимости от среды, в которой они существуют. Как таковые, исследователи машинного поведения должны сосредоточиться на характеристике поведения агентов в различных средах, подобно тому, как ученые-бихевиористы стремятся охарактеризовать политическое поведение в различных демографических и институциональных контекстах.

В-третьих, машины демонстрируют поведение, принципиально отличное от поведения животных и людей, поэтому мы должны избегать чрезмерного антропоморфизма и зооморфизма. Даже если заимствование существующих поведенческих научных методов может оказаться полезным для изучения машин, машины могут проявлять формы интеллекта и поведения, качественно отличные, даже чуждые, от тех, которые наблюдаются в биологических агентах. Кроме того, ученые ИИ могут анализировать и модифицировать системы ИИ легче и тщательнее, чем это происходит со многими живыми системами. Хотя параллели существуют, изучение систем ИИ обязательно будет отличаться от изучения живых систем.

В-четвертых, изучение поведения машин потребует междисциплинарных усилий и повлечет за собой все проблемы, связанные с такими исследованиями. Решение этих проблем имеет жизненно важное значение. Университеты и государственные финансовые учреждения могут играть важную роль в разработке крупномасштабных, нейтральных и надежных междисциплинарных исследований 141.

В-пятых, изучение поведения машин часто требует экспериментального вмешательства для изучения взаимодействия человека и машины в реальных условиях.Эти мероприятия могут изменить общее поведение системы, возможно, оказывая негативное воздействие на обычных пользователей 144. Такие этические соображения требуют тщательного надзора и стандартизированных рамок.

Наконец, изучение интеллектуальных алгоритмических или роботизированных систем может привести к юридическим и этическим проблемам для исследователей, изучающих поведение машин. Алгоритмы обратного проектирования могут потребовать нарушения условий обслуживания некоторых платформ, например, при настройке поддельных персонажей или маскировке истинных идентичностей. Создатели или сопровождающие системы интересов могут втянуть исследователей в юридические проблемы, если исследование наносит ущерб репутации их платформ. Кроме того, остается неясным, может ли нарушение условий службы подвергнуть исследователей гражданскому или уголовному наказанию (например, через закон о компьютерном мошенничестве и злоупотреблениях в США), что может еще больше лишить этот тип исследований 145.

Понимание поведения и свойств агентов ИИ-и воздействия, которое они могут оказать на человеческие системы-имеет решающее значение. Общество может извлечь огромную пользу из новой эффективности и улучшения процесса принятия решений, которые могут быть получены от этих агентов. В то же время эти преимущества могут ослабнуть без минимизации потенциальных ловушек включения агентов ИИ в повседневную жизнь человека.

Автор вклад

И. Р., М. К. и Н.О. задумали идею, произвели цифры и составили рукопись. Все авторы внесли свой вклад в содержание, доработали и отредактировали рукопись.

соперничающие интересы

Авторы не заявляют о каких-либо конкурирующих интересах.

Выражение благодарности

Авторы выражают признательность следующим содержанием: Р. И. от этики и управления, искусственного интеллекта Фонда; Ж. Б. от НФС награды вдохновляют-1344227 и бигдата-1447634, DARPA в непрерывное обучение машин программу, АРО договора W911NF-16-1-0304 Ж. Ф. Б от АНР-Labex последней; н. А. с Пионером грант от фонда Роберта Вуда Джонсона Фонда; И. Д. С. из НФС (ИОС-1355061), с ОНР (N00014-09-1-1074 и N00014-14-1-0635), в АРО (W911NG-11-1-0385 и W911NF14-1-0431), в структур - УНД Innovationsfunds für умереть состояние земли Баден-Вюртемберг, Общества имени Макса Планка, германское научно-исследовательское общество и Центр передового опыта 2117 “центр углубленным изучением коллективного поведения” (код: 422037984); Д. Л. от Управления директора Национальной разведки (одни), разведки передовой исследовательской деятельности проектов (целью) в соответствии с договором 2017-17061500006; J. B. T из центра мозга, сознания и машин (CBMM) под NSF STC award CCF – 1231216; M. W из Института Будущего жизни.

Мнения и выводы, содержащиеся в настоящем документе, являются мнениями авторов и не должны рассматриваться как обязательно представляют официальную политику, явно выраженных или подразумеваемых, одни, целью или правительства США. Правительство США уполномочено воспроизводить и распространять перепечатки для государственных целей, независимо от любых авторских аннотаций в них.