Этические дилеммы развития искусственного интеллекта в цифровую эпоху – современные дискуссии и полемика

ОСНОВНЫЕ ТЕЗИСЫ

Гаспарян Диана Эдиковна

Публичное восприятие ИИ. В целом восприятие потенциала и возможных эффектов ИИ как широкой публикой, так и экспертным сообществом противоречиво. Существуют оптимистические и пессимистические тенденции, хотя следует отметить, что первые в общем преобладают. Оптимизм и пессимизм часто связаны с разными аспектами и областями применения ИИ. Так, широко цитируемый анализ Фаста и Хорвица (Fast & Horvitz 2017) свидетельствует, что общественное обсуждение ИИ резко возросло с 2009 года. Характер общественных дискуссий неизменно был более оптимистичным, чем пессимистичным. Однако Фаст и Хорвиц обнаружили, что в последние годы усилились опасения по поводу потери контроля над ИИ, этических проблем ИИ и негативного влияния ИИ на занятость. Однако надежды на ИИ в здравоохранении и образовании со временем возросли.

Мюллер и Бустрём (Müller & Bostrom 2016) отмечают беспокойство в некоторых экспертных кругах по поводу того, что через несколько десятилетий появится высокоуровневый машинный интеллект и сверхразумный ИИ, которые принесут с собой значительные риски для человечества. В других кругах эти проблемы игнорируются или считаются научной фантастикой. Мюллер и Бустрём распространили анкету среди четырех групп экспертов в 2012/2013 гг. По средней оценке респондентов, вероятность того, что высокоуровневый искусственный интеллект будет разработан примерно в 2040–2050 годах, составляет один к двум, а к 2075 году вероятность возрастет до девяти из десяти. Эксперты ожидают, что интеллектуальные системы перейдут к сверхинтеллекту (superintelligence) менее чем через 30 лет после разработки высокоуровневого машинного интеллекта. По их оценкам, вероятность того, что это развитие

окажется «плохим» или «чрезвычайно плохим» для человечества, составляет примерно один к трем.

В 2018 году аналитический центр Pew Research подготовил доклад, основанный на опросе «979 технологических пионеров, инноваторов, разработчиков, лидеров в области бизнеса и политики, исследователей и активистов». Всем участникам были заданы следующие вопросы: «Считаете ли вы, что к 2030 году развитие ИИ усилит человеческий потенциал и расширит его возможности? Будет ли большинство людей, благодаря ИИ, жить лучше, чем сегодня? Или, скорее всего, развитие ИИ и связанных с ним технологических систем уменьшит человеческую автономию и свободу действий до такой степени, что большинство людей не станут жить лучше?» В целом, несмотря на опасения, 63% респондентов выразили надежду на то, что большинство людей в 2030 году в основном будут жить лучше благодаря ИИ, а 37% заявили об обратном (Anderson et al. 2018; 4).

Составители доклада, проанализировав замечания экспертов, сгруппировали их опасения в пять основных тем:

1. Человеческая деятельность: люди теряют контроль над своей жизнью. Принятие решений по ключевым аспектам автоматически передается управляемым кодом инструментам «черного ящика». Людям не хватает информации, и они не знают, как работают инструменты. Этот эффект будет усиливаться по мере того, как автоматизированные системы будут становиться все более распространенными и сложными.

2. Злоупотребление данными: использование и сбор данных в сложных системах предназначены для получения прибыли.

Ценности и этика часто не встроены в цифровые системы, принимающие решения за людей. Эти системы объединены в глобальную сеть, и их нелегко регулировать.

3. Потеря работы: захват рабочих мест со стороны ИИ увеличит экономический разрыв, что приведет к социальным потрясениям. Эффективность и другие экономические преимущества машинного интеллекта,

основанного на коде, будут продолжать создавать неурядицу во всех аспектах человеческого труда. В то время как одни опрошенные ожидают появления новых рабочих мест, другие обеспокоены массовой потерей рабочих мест, расширением экономического неравенства.

4. Укоренение зависимости: снижение когнитивных, социальных навыков и навыков выживания. Многие эксперты считают, что ИИ расширяет человеческие возможности, но некоторые предсказывают обратное: растущая зависимость людей от управляемых машинами сетей подорвет их способность думать самостоятельно и действовать независимо от автоматизированных систем и эффективно взаимодействовать с другими людьми.

Костас Александридис, автор книги «Изучение сложной динамики в многоагентных интеллектуальных системах», предсказывает: «Многие из наших повседневных решений будут автоматизированы с минимальным вмешательством со стороны конечного пользователя. Автономия и/или независимость будут принесены в жертву и заменены удобством. Новые поколения граждан будут все больше и больше зависеть от сетевых структур и процессов ИИ. Существуют проблемы, которые необходимо решать с точки зрения критического мышления и неоднородности. Сетевая взаимозависимость, скорее всего, повысит нашу уязвимость к кибератакам. Существует также реальная вероятность того, что будут более четкие различия между цифровыми «имущими» и «неимущими», а также между зависящими от технологий цифровыми инфраструктурами. Наконец, возникает вопрос о новых «командных высотах» владения и контроля над цифровой сетевой инфраструктурой» (Anderson et al. 2018; 8).

Социетальное влияние и зависимость от технологий ИИ. Авторы анализируемой литературы отмечают глубокое воздействие, которое технологии ИИ, по их оценкам, окажут на человеческие сообщества, повседневную жизнь и личную автономию по мере развития таких технологий. Среди монографий и сборников статей на эти темы следует выделить Müller 2015, Yampolskiy 2015, Elliot 2018, Elliot 2021. Так, Эллиот (Elliot 2018) утверждает, что в первую очередь

необходимо рассматривать «ИИ-революцию» не с точки зрения предполагаемого появления общего ИИ или наделённых ИИ роботов, а в аспекте глубоких изменений в нашей повседневной жизни, которые вызывают проникновение в неё «умных» технологий — от получения рекомендаций на Amazon до запроса Uber, от получения информации от виртуальных личных помощников до общения с чат-ботами.

Важный аспект социетального влияния ИИ — так называемая петля обратной связи (feedback loop). Идеи, предубеждения, стереотипы и предвзятое поведение, существующие в массивах, данных, влияют на решения ИИ, которые потом (без ведома пользователя) возвращаются в человеческое сознание и подкрепляют существующие установки. Таким образом, отношение человек-ИИ представляет собой систему обратной связи с подкрепляющими друг друга стимулами.

В качестве примера ситуации, когда оптимистические ожидания от ИИ накладываются на тревогу о замещении алгоритмами человеческих работников, приведём исследование Парк и коллег (Park et al. 2021). Студенты-медики из США (n=156) согласились (75%) с тем, что ИИ сыграет важную роль в будущем медицины. Большинство (66%) согласилось с тем, что диагностическая радиология из-за ИИ изменится больше других областей. Почти половина (44%) сообщили, что искусственный интеллект уменьшил их интерес к радиологии в качестве будущей области специализации. При этом большинство опрошенных студентов отметили, что черпают свои представления об ИИ из популярной прессы. Таким образом, освещение и репутация ИИ в прессе служит примером ещё одной петли обратной связи по отношению к перспективам реального ИИ: чем больше студенты читают материалы об ИИ, тем меньше хотят заниматься радиологией и тем больше возникает потребность во внедрении ИИ в эту область медицины и диагностики.

В качестве возможного эффекта влияния ИИ на самосознания и одновременно роста зависимости приведём пример Веллора. Веллор (Vallor 2015) использует социо-экономический термин deskilling, означающий устаревание

определённых трудовых навыков благодаря автоматизации. Термин вводится, чтобы поднять проблему морального разучения (moral deskilling). Веллор рассматривает моральное разучение на примерах автоматизированного оружия, практик в новых медиа и социальной робототехники (использование роботов-сиделок и т.д.). Веллор заключает, что моральный дескиллинг повторяет фундаментальную амбивалентность социо-экономического дескиллинга. Передача опредленных навыков машинам позволяет нам либо забыть важные навыки (deskilling), в том числе моральные, либо позволяет освоить новые (reskilling), в том числе те, которые по разным причинам являются для нас более предпочтительными (upskilling). Поскольку моральные навыки являются важными предпосылками для эффективного развития практической мудрости и добродетельного характера, Веллор заключает, что феномен морального разучения требует пристального внимания этиков и философов технологии. Делается вывод, что понимание неоднозначного и открытого горизонта влияния новых технологий на моральные навыки может помочь нам не только лучше определить моральные риски и преимущества таких технологий, но и предвидеть новые возможности разработки и проектирования, которые конкретным образом реализуют потенциальные преимущества этих технологий, одновременно сводя к минимуму вред, который они в противном случае могли бы нанести нашим собственным моральным способностям.

Вместе с тем неясна сама желательность дескиллинга и границы делегирования интеллектуальным системам принятия моральных и иных решений. Поскольку принятие решений человеком улучшается за счет факторов, отличных от интеллекта — таких, как мотивационные процессы, эмоции и опыт (Innes & Morrison 2021), неясно, может и должен ли алгоритм ИИ обладать аналогичными свойствами и пользоваться ими в ходе принятия решений.

Лояльность и доверие. В целом в анализируемой литературе утверждают, что доверие является центральным компонентом взаимодействия между людьми и ИИ. Низкий уровень доверия может привести к неправильному использованию, злоупотреблению или отказу от использования определённой технологии.

Широко цитируемое исследование Люо и коллег (Luo et al. 2019) отталкивается от данных эксперимента на более, чем 6200 клиентах, которые получали тщательно структурированные коммерческие звонки либо от чат-ботов, либо от сотрудников-людей (кто звонил кому определялось случайным образом). Результаты показали, что с точки зрения продаж скрытые чат-боты были так же эффективны, как опытные работники, и в четыре раза эффективнее, чем неопытные работники. Однако раскрытие чат-бота до разговора между машиной и покупателем снижает количество покупок более чем на 79,7%. Раскрытие чат-бота также существенно сокращает продолжительность звонка. Когда клиенты знают, что собеседник не человек, они отвечают скупо и покупают меньше, потому что считают бота менее осведомленным и менее чутким. Негативный эффект раскрытия информации, по-видимому, обусловлен субъективным отношением человека к машинам, несмотря на объективную компетентность чат-ботов с искусственным интеллектом.

Авторы анализируемой литературы утверждают, что с публичным доверием и лояльностью к ИИ тесно связана проблема объяснимости ИИ. За последнее десятилетие, благодаря наличию больших наборов данных и большей вычислительной мощности, системы машинного обучения достигли (сверх)человеческой производительности в самых разных задачах. Примеры такого быстрого развития можно увидеть в распознавании изображений, анализе речи, стратегическом планировании игр и многом другом. Проблема многих современных моделей заключается в отсутствии прозрачности и интерпретируемости. Непрозрачность и непонятность являются серьезными недостатками во многих приложениях ИИ, например в здравоохранении и финансах, где обоснование решения модели ИИ выступает требованием для появления доверия к модели. В свете этих проблем предметом интереса исследовательского сообщества стал концепт объяснимого искусственного интеллекта (explainable artificial intelligence, XAI) (Došilović 2018).

Юдон и коллеги (Hudon et al. 2021) концептуализируют объяснимость процессов принятия решения в ИИ. Объяснимый искусственный интеллект (XAI)

призван обеспечить прозрачность систем ИИ за счет перевода, упрощения и визуализации его решений для пользователя. Хотя общество по-прежнему скептически относится к системам ИИ, исследования показывают, что прозрачные и объяснимые системы ИИ приводят к повышению доверия между людьми и ИИ. Результаты Юдона и коллег демонстрируют, что как порядок презентации схем принятия решений алгоритмом, так и морфологическая ясность влияют на когнитивную нагрузку пользователей. Кроме того, была выявлена отрицательная корреляция между когнитивной нагрузкой и доверием к системе ИИ.

Джейкови и коллеги (Jacove et al. 2021) предлагают модель доверия, вдохновленную межличностным доверием (между людьми), как его определяют социологи. Их модель опирается на два ключевых свойства: уязвимость пользователя и способность предвидеть эффект решений, принимаемых ИИ. Авторы формализуют «договорное доверие», то есть такое доверие между пользователем и ИИ, при котором будет соблюдаться некий неявный или явный договор, «надежность» (trustworthiness) и понятия «оправданного» и «неоправданного» доверия. Авторы обсуждают вопрос, как разработать заслуживающий доверия ИИ, и как оценить, проявилось ли доверие и оправдано ли оно.

Джейкеш и коллеги (Jakesch et al. 2019) утверждают, что мы вступаем в эру общения, опосредованного ИИ (AI-mediated communication, AI-MC), когда межличностное общение не просто опосредуется технологиями, но и оптимизируется, дополняется или генерируется искусственным интеллектом. Авторы впервые рассмотрели потенциальное влияние AI-MC на саморепрезентацию пользователей Интернета. Когда участники думали, что видят смешанный набор профилей (и созданные ИИ, и людьми), они не доверяли хозяевам жилья на Airbnb, чьи профили были помечены как созданные ИИ. Таким образом, в выборке, содержащей и человеческие, и "машинные" профили, первые пользовались большим доверием.

Феррари и Луа (Ferrario & Loi 2022) предлагают философское объяснение взаимосвязи между объяснимостью ИИ и доверием к ИИ. Это объяснение связывает обоснование надежности ИИ с необходимостью контроля за ним во время его использования. Они предполагают, что объяснимость способствует доверию к ИИ тогда и только тогда, когда она способствует обоснованному и гарантированному парадигматическому доверию к ИИ — то есть доверию при наличии обоснованного убеждения в том, что ИИ заслуживает доверия. Такое доверие, в свою очередь, каузально способствует тому, что на ИИ можно положиться при отсутствии наблюдения. Авторы различают доверие индивидуального пользователя к ИИ, где объяснимость не способствует доверию. Напротив, при рассмотрении общественного доверия к ИИ, используемому человеком, они утверждают, что объяснимость может способствовать доверию. Такой подход объясняет кажущийся парадокс в отношениях ИИ и человека: чтобы доверять ИИ, мы должны позволить индивидуальным пользователям ИИ не доверять ИИ полностью.

В обзоре Гликсон и Вулли (Glickson & Woolley 2020) отобраны существующие эмпирические исследования детерминант человеческого доверия к ИИ, проведенные в различных дисциплинах за последние 20 лет. Авторы отталкиваются от трёх форм ИИ (роботический, виртуальный и встроенный ИИ) и уровня машинного интеллекта как важных предпосылок для определения человеческого доверия к алгоритмам. Обзор соотносит три формы ИИ с тремя вариантами машинного поведения: 1) присутствием (tangibility), 2) антропоморфностью и 3) сходством поведения ИИ с поведением человека (immediacy behaviors).

1) Согласно обзору Гликсон и Вулли, физическое присутствие робота может усилить симпатию, но также и вызвать страх. Присутствие виртуальной ИИ-«личности» повышает симпатию и эмоциональное доверие к ИИ. Что касается последней формы, то неосведомленность о наличии встроенного в продукт ИИ может вызвать гнев пользователя. Положительные эмоции может вызвать хорошая репутация фирмы-производителя продукта.

2) Антропоморфная внешность робота в целом усиливает положительные эмоции взаимодействующего человека, но может и вызывать дискомфорт. Антропоморфность виртуального ИИ в основном повышает доверие, но также создает большие ожидания в плане способностей ИИ. Привлекательная внешность и отдельные факторы, вроде этнической принадлежности или сходства лица виртуального ИИ с лицом пользователя, повышают доверие.

3) Человекообразное поведение робота создаёт у человека высокое эмоциональное доверие. Роботы, совершающие ошибки, вызывают большую симпатию, чем роботы с «безупречным» поведением. Что касается виртуальных ИИ, человекоподобное поведение повышает доверие и симпатию пользователя, но эффект зависит от изначальных установок пользователя.

В целом, согласно анализируемой литературе, можно сделать предварительный вывод о взаимосвязи динамики доверия, лояльности и зависимости в отношении ИИ. Однако нам не удалось найти исследования, которые бы рассматривали все три аспекта одновременно.

Ценности. Большой объём литературы посвящён теме ценностного согласования ИИ с «человеческими» ценностями (AI value alignment). Цель такого согласования — обеспечить корректное соответствие ИИ человеческим ценностям (Russell 2019, 137). Согласованность ценностей — это свойство интеллектуального агента преследовать только те цели, которые выгодны людям. Успешное согласование ценностей должно гарантировать, что общий искусственный интеллект не сможет преднамеренно или непреднамеренно выполнять действия, которые отрицательно влияют на людей.

Многие авторы предлагают ограничивать поведение интеллектуальных систем «машинной этикой» (machine ethics), чтобы обеспечить положительные социальные результаты разработки таких систем. Брандидж (Brundage 2015) утверждает, что хотя машинная этика может повысить вероятность этичного поведения в некоторых ситуациях, она не может гарантировать этичного поведения из-за природы этики, ограничений вычислительных агентов и сложности мира. Кроме того, машинная этика, даже если бы она была «решена»

на техническом уровне, была бы недостаточной для обеспечения положительных социальных результатов применения систем ИИ.

Ридл и Харрисон (Riedl & Harrison 2016) утверждают, что для успешного согласования ценностей ИИ должен сам изучать ценности. Они считают, что искусственный интеллект, способный читать и понимать истории, может изучать ценности, негласно поддерживаемые порождающие истории культурой. Истории могут использоваться инженерами ИИ для создания ориентированного на ценности сигнала вознаграждения с подкреплением для обучающегося ИИ с целью предотвращения квази-психотического поведения со стороны ИИ.

Гэбриел во влиятельной статье (Gabriel 2020) доказывает, что методы машинного обучения, которые мы используем для ценностного согласования, и ценности, с которыми происходит согласование, не полностью независимы друг от друга. То, как мы проектируем ИИ, влияет на ценности, которые мы можем ему привить, и более четкое понимание ценностного измерения может продуктивно направить исследования в области ИИ. Как следствие, в проектировании ИИ невозможно полностью «вынести за скобки» нормативные вопросы. Вместо этого они должны стать частью объединенной исследовательской программы. Гэбриел утверждает, что мы не должны стремиться согласовать ценности ИИ только с инструкциями, выраженными намерениями или выявленными предпочтениями. Правильно согласованный ИИ должен будет учитывать различные формы неэтичного или неразумного поведения и строиться на инженерных принципах, которые предотвращают такое поведение. Один из способов сделать это — встроить объективные ограничения того, что могут делать искусственные агенты, заключает Гэбриел. Еще более полезным был бы набор широко поддерживаемых людьми принципов, несмотря на существование различных систем убеждений. Потребуется работа как с точки зрения технического уточнения понятий, которые фигурируют в принципах, так и с точки зрения определения правильных принципов.

Гэбриел утверждает, что основная задача ценностного согласования состоит не в том, чтобы определить истинную моральную теорию и затем на её основе

запрограммировать ИИ, а в том, чтобы определить принципы ИИ, которые будут широко считаться справедливыми. Проблема согласования в этом смысле политическая, а не метафизическая. Чтобы решить эту проблему, Гэбриел рекомендует искать такие принципы, которые будут поддерживаться глобальным перекрывающимся консенсусом мнений. Эти принципы должны быть выбраны за завесой невежества или утверждены посредством демократических процессов.

Помимо определения содержания принципов ценностного согласования, мы должны также подумать о том, как можно осуществить процесс интеграции и поиска консенсуса. В идеале процесс, используемый для определения принципов согласования ИИ, должен быть процедурно справедливым в том смысле, что он не дает произвольного преимущества одной стороне; конкретным в том смысле, что он дает подробное руководство; стабильным и надежным, ведущим к принципам, которые можно поддерживать с течением времени; всеобъемлющим, оставляющим мало пробелов; и действительно инклюзивным, включающим аргументированные мнения всех, кто участвует в отборе принципов. Еще одним важным качеством процесса отбора является его способность справляться с возможностью широко распространенной моральной ошибки. Учитывая, что мы можем совершать ошибки такого рода, было бы неправильным слишком сильно привязывать ИИ к морали текущей эпохи.

Намечающие усилия в области ценностного согласования и «этической разработки» ИИ (ethical design) критически исследованы Грин (Greene et al. 2019) и коллегами. Авторы предлагают обзор ценностных деклараций влиятельных организаций по изучению ИИ методами дискурс-анализа. Грин и коллеги обнаруживают семь невысказанных допущений, которые разделяют все исследованные ими заявления о ценностях разработчиков ИИ: вера в объективно измеряемые универсальные проблемы, экспертный надзор, направляемый ценностями детерминизм, разработку как объект этической проверки, ответственное строительство, легитимность, определяемую заинтересованными сторонами и то, что авторы называют «машинным переводом» - тенденцию

разработчиков ИИ приписывать алгоритмическим инструментам способность автономно решать моральные проблемы (ср. с Vallor 2015).

Лагерквист (Lagerkvist 2020) утверждает, что нынешний момент технологической трансформации и эскалации многогранных и взаимосвязанных глобальных кризисов представляет собой "цифровую предельную ситуацию". Она называет достижение постдисциплинарной, интегративной и генеративной формы гуманитарных и социальных наук "методом надежды", который вовлекает разработчиков ИИ в поиск инклюзивного и открытого будущего в рамках экзистенциальной и экологической устойчивости.

Учитывая быстрое распространение ИИ, вопросы влияния ИИ на человеческий мир уже актуальны. Однако ИИ всё ещё молодая технология (или, скорее, большое семейство технологий), долгосрочные социетальные и психологические эффекты которой пока не ясны. Поэтому инженеры ИИ, по мнению исследователей ИИ в социальных и гуманитарных дисциплинах, должны постоянно иметь в виду, как должна выглядеть эта технология с точки зрения прогнозируемых этических, психологических, политических и социальных её последствий. Согласно рассмотренной литературе, следует трактовать ИИ как технологию, находящуюся в обратной связи с её пользователями, которая не только испытывает влияние человеческих агентов (инженеров, пользователей, экспертных комиссий), но в свою очередь уже оказывает и продолжит оказывать существенное и не вполне прозрачное влияние на их сознание и поведение.

Литература

1. Anderson, J.; Rainie, L. & Luchsinger, A. (2018). Artificial Intelligence and the Future of Humans. Pew Research Center, December, 2018. Retrieved from https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/12/PI_2018.12.10_future-of-ai_FINAL1.pdf

2. Brundage, M. Limitations and Risks of Machine Ethics, in Risks of Artificial Intelligence. Edited by Vincent C. Müller. Chapman and Hall/CRC, 2015.

3. Devillers, L., Fogelman-Soulié, F., Baeza-Yates, R. (2021). AI & Human Values. In: Braunschweig, B., Ghallab, M. (eds) Reflections on Artificial Intelligence for Humanity. Lecture Notes in Computer Science(), vol 12600. Springer, Cham. https://doi.org/10.1007/978-3-030-69128-8_6

4. Došilović F. K.; Brčić M.; Hlupić N. (2018). "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.

5. Elliott, A. (2018). The Culture of AI: Everyday Life and the Digital Revolution (1st ed.). Routledge. https://doi.org/10.4324/9781315387185 (https://www.taylorfrancis.com/books/mono/10.4324/9781315387185/culture-ai-anthony-elliott)

6. Elliott, A.(ed.) (2021). The Routledge Social Science Handbook of AI. London: Routledge, 2021. https://doi.org/10.4324/9780429198533 (https://www.taylorfrancis.com/books/edit/10.4324/9780429198533/routledge-social-science-handbook-ai-anthony-elliott)

7. Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10635

8. Ferrario, A. and Loi, M. (2022). How Explainability Contributes to Trust in AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

9. Gabriel, I. Artificial Intelligence, Values, and Alignment. *Minds & Machines* 30, 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2. Retrieved from https://link.springer.com/article/10.1007/s11023-020-09539-2

10. Glikson, Ella; Woolley, Anita Williams (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, 2018,

Vol. 14, No. 2. https://doi.org/10.5465/annals.2018.0057. Retrieved from https://journals.aom.org/doi/full/10.5465/annals.2018.0057

11. Greene, D.; Hoffmann, A. L.; Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In Tung Bui, editor, 52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019, 1-10.

12. Hudon, A., Demazure, T., Karran, A., Léger, PM., Sénécal, S. (2021). Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. In: Davis, F.D., Riedl, R., vom Brocke, J., Léger, PM., Randolph, A.B., Müller-Putz, G. (eds) Information Systems and Neuroscience. NeuroIS 2021. Lecture Notes in Information Systems and Organisation, vol 52. Springer, Cham. https://doi.org/10.1007/978-3-030-88900-5_27

13. Innes, J. M.; Morrison, B. W. (2021). Artificial intelligence and psychology, in Elliott, A. (ed.). The Routledge Social Science Handbook of AI. London: Routledge, 2021.

14. Jacovi, A.; Marasović, A.; Miller, T.; Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

15. Jakesch, M.; French, M.; Xiao, M.; Hancock, J. T.; and Naaman M. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 239, 1–13. https://doi.org/10.1145/3290605.3300469

16. Lagerkvist, Amanda. (2020). Digital Limit Situations: Anticipatory Media Beyond ‘The New AI Era’. Journal of Digital Social Research. 2. 10.33621/jdsr.v2i3.55.

17. Luo X.; Tong S.; Fang Z.; Qu Z. (2019) Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. Marketing Science 38(6):937-947. https://doi.org/10.1287/mksc.2019.1192

18. Müller, V. C. (ed.). Risks of Artificial Intelligence. Chapman and Hall/CRC, 2015. (https://www.taylorfrancis.com/books/edit/10.1201/b19187/risks-artificial-intelligence-vincent-m%C3%BCller)

19. Müller, V .C., Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller, V.C. (eds) Fundamental Issues of Artificial Intelligence. Synthese Library, vol 376. Springer, Cham. https://doi.org/10.1007/978-3-319-26485-1_33

20. Park, C. J., Paul H. Yi, Eliot L. Siegel, Medical Student Perspectives on the Impact of Artificial Intelligence on the Practice of Medicine, Current Problems in Diagnostic Radiology, Volume 50, Issue 5, 2021, Pages 614-619, ISSN 0363-0188, https://doi.org/10.1067/j.cpradiol.2020.06.011.(https://www.sciencedirect.com/science/article/pii/S0363018820301249)

21. Riedl, M., & Harrison, B. (2016). Using Stories to Teach Human Values to Artificial Agents. In AAAI Workshops. Retrieved from https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12624/12352

22. Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Bristol: Allen Lane.

23. Vallor, S. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. Philos. Technol. 28, 107–124 (2015). https://doi.org/10.1007/s13347-014-0156-9 Retrieved from https://link.springer.com/article/10.1007/s13347-014-0156-9

24. Weber, J.; Prietl, B. (2021). AI in the age of technoscience. On the rise of data-driven AI and its epistem-ontological foundations, in Elliott, A.(ed.) (2021). The Routledge Social Science Handbook of AI. London: Routledge, 2021. https://doi.org/10.4324/9780429198533

25. Yampolskiy, R.V. (2015). Artificial Superintelligence: A Futuristic Approach (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18612 (https://www.taylorfrancis.com/books/mono/10.1201/b18612/artificial-superintelligence-roman-yampolskiy).