Научная статья
УДК 34:004:340.1:004.8
EDN: https://elibrary.ru/pltfwo
DOI: https://doi.org/10.21202/jdtl.2025.17

Агентный искусственный интеллект: правовые и этические вызовы автономных систем

Гордон Боуэн

Университет Англия Рёскин, Кембридж, Великобритания

Ключевые слова

автономность,
агентный искусственный интеллект,
искусственный интеллект,
ответственность,
право,
правовое регулирование,
программирование,
риск,
цифровые технологии,
этика

Аннотация

Цель: определить специфические правовые и этические проблемы агентного искусственного интеллекта и выработать рекомендации по созданию защитных механизмов для обеспечения ответственного функционирования автономных ИИ-систем.

Методы: исследование носит концептуальный характер и основано на системном анализе научной литературы по вопросам этики искусственного интеллекта, правового регулирования автономных систем и социального взаимодействия ИИ-агентов. В работе применяются сравнительный анализ различных типов ИИ-систем, исследование потенциальных рисков и преимуществ агентного искусственного интеллекта, а также междисциплинарный подход, интегрирующий достижения в сфере права, этики и компьютерных наук для формирования комплексного понимания проблематики.

Результаты: установлено, что агентный искусственный интеллект, обладая автономностью принятия решений и способностью к социальному взаимодействию, создает качественно новые правовые и этические вызовы по сравнению с традиционными ИИ-ассистентами. Выявлены основные категории потенциального вреда: прямое воздействие на пользователей через открытые и скрытые действия, манипулятивное влияние на поведение и кумулятивный вред от длительного взаимодействия. Определена необходимость распределения ответственности между тремя ключевыми субъектами: пользователем, разработчиком и владельцем системы агентного искусственного интеллекта.

Научная новизна: впервые проведен системный анализ этических аспектов агентного искусственного интеллекта как качественно нового класса автономных систем, отличающихся от традиционных ИИ-ассистентов степенью независимости и социальной интерактивности. Разработана типология потенциальных рисков социального взаимодействия с агентными интеллектуальными системами, и предложена концептуальная модель распределения правовой и этической ответственности в триаде «пользователь – разработчик – владелец».

Практическая значимость: результаты исследования формируют теоретическую основу для разработки этических принципов и правовых норм регулирования агентного искусственного интеллекта в условиях растущего рынка автономных интеллектуальных систем. Полученные выводы могут быть использованы законодателями при создании нормативной базы, разработчиками при проектировании защитных механизмов, а также организациями при внедрении агентных систем искусственного интеллекта в различных сферах экономической деятельности.

Для цитирования

Боуэн, Г. (2025). Агентный искусственный интеллект: правовые и этические вызовы автономных систем. *Journal of Digital Technologies and Law*, *3*(3), 431–445. https://doi.org/10.21202/jdtl.2025.17

Содержание

Введение
1. Обзор литературы
   1.1. Три типа ИИ-агентов
   1.2. Социальное взаимодействие ИИ-агентов
   1.3. Принятие решений с помощью агентного ИИ
2. Практические рекомендации
Заключение
Список литературы

Введение

В настоящее время мы наблюдаем расширение возможностей ИИ-агентов, таких как коммуникационные навыки и сложные рассуждения без вмешательства человека. ИИ-ассистенты привязаны к пользователям, но агенты с искусственным интеллектом имеют определенную степень свободы[1]. Глобальная рыночная стоимость агентного ИИ в 2024 г. составляла 5,1 млрд долл. США и, как ожидается, к 2030 г. увеличится до 47 млрд долл. США при совокупном годовом темпе роста в 44 %[2]. Степень свободы, которой в настоящее время обладают агенты с искусственным интеллектом, требует правовых и этических рамок для регулирования их поведения. Как владелец/разработчик агентного ИИ, так и соответствующее программное обеспечение нуждаются в мониторинге с юридической и этической точек зрения. Но, чтобы не потерять преимущества агентного ИИ, необходимо соблюдать баланс. Для «статичных» помощников с ИИ уже выработаны этические и правовые рамки, которые контролируются соответствующими действиями или привязаны к ним. Однако они требуют пересмотра по отношению к ИИ-агентам. Как же именно следует изменить эти рамки для агентов с ИИ? Требуется ли агентному ИИ

1 Morris, B. (2024). Beyond Intelligence: The Impact of Advanced AI Agents. https://clck.ru/3NedB2

2 Vailshery, L. S. (2025). Global market value of agentic AI 2030. https://clck.ru/3NedDe

общественное сознание, чтобы ориентироваться в новой этической и правовой среде? Главная цель нашего исследования – определить, как будут развиваться дебаты об этике и правовом поле для агентов с искусственным интеллектом. Статья состоит из введения, обзора литературы, основной части и выводов.

1. Обзор литературы

1.1. Три типа ИИ агентов

ИИ-агенты называют также сложными системами ИИ; в настоящее время эта область исследований бурно развивается (Kapoor et al., 2024). Сложные системы ИИ – это лучший способ использовать и максимизировать модели ИИ и, возможно, один из важнейших трендов 2024 г.[3] Сложные системы искусственного интеллекта во многих отношениях отличаются от обычных ИИ-систем (больших языковых моделей). Так, они решают более сложные задачи, чаще используются в реальных условиях и могут решать проблемы, на которые нет однозначного ответа; для них могут потребоваться пользовательские интерфейсы агент-компьютер (Yang, Jimenez et al., 2024). Сложная ИИ-система может управлять несколькими агентами, что сказывается на ее стоимости (Kapoor et al., 2024).

В отношении традиционного ИИ считается, что агенты способны воспринимать окружающую среду и воздействовать на нее (Russell & Norvig, 1995); с этой точки зрения термостат может быть классифицирован как агент (Kapoor et al., 2024). Агентные ИИ-системы часто рассматриваются как различные системы искусственного интеллекта с той или иной степенью агентной способности[4]. Существует три типа ИИ-агентов (Alberts, Keeling et al., 2024): артефакты (агент интерпретирует данные в социальной среде), интерактивные системы (ведут себя как социальные субъекты) и диалоговые агенты (выполняют социальные роли). ИИ-агенты как социальные субъекты должны уметь общаться и, таким образом, взаимодействовать с пользователем, но это означает нечто большее, чем просто быть приятными, дружелюбными, говорить правду и не использовать ненормативную лексику. Ожидается, что взаимодействие будет контекстуальным, что требует понимания личности пользователя, социальной среды и ситуационного контекста. Это приведет к тому, что ИИ-агенты будут выдавать информацию без запроса, а значит, вносить собственные предложения (Alberts, Keeling et al., 2024).

1.2. Социальное взаимодействие ИИ-агентов

Технологии можно считать социальными, поскольку они внедрены в социальную среду и используются в ней. Это предположение разделяют исследователи, изучающие идеологический аспект технологий, в частности, культурные предубеждения и ценности, которые заложены в используемых технологиях (Bender et al., 2021; Shelby et al., 2023). Системы, которые не являются социально интерактивными, рассматриваются как вредные (Alberts, Keeling et al., 2024). Примером вреда «пассивной»

---

[3] Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., & Ghodsi, A. (2024). The Shift from Models to Compound AI Systems. https://clck.ru/3NedKd

[4] Ng, A. (2024). Welcoming Diverse Approaches Keeps Machine Learning Strong. https://clck.ru/3NedNZ

технологической системы служат данные об уровне обученности, которые искажают демографическую картину (Bender et al., 2021).

Помимо пассивных технологий, интерактивные технологии имеют определенное назначение и являются ценными для социума (Grimes et al., 2021). Кроме того, локальные действия системы интерпретируются с точки зрения человека и общества. На этом принципе основана философия исследования «Компьютеры как социальные субъекты», согласно которому люди, взаимодействуя с компьютерными системами и технологиями, применяют человеческие социальные нормы и ожидания (Nass et al., 1994).

Интерактивные системы имитируют поведение или качества человека. Это достигается с помощью социальных сигналов (говорение от первого лица и выражение эмоций) (Grimes et al., 2021). Диалоговые агенты способны давать адекватную реакцию в знакомых социальных ситуациях. Система ИИ может выполнять социальную роль, в которой ИИ-агент является другом или терапевтом. Однако ИИ-агент может также сказать что-то оскорбительное и расстроить пользователя, либо проявить фамильярность, что ассоциируется с презрением или бесчувственностью (Alberts, Keeling et al., 2024). Социальные взаимодействия, которые могут причинить вред, классифицируются следующим образом (Alberts, Keeling et al., 2024):

– прямой вред пользователю – побуждение к открытым действиям, таким как оскорбление из-за используемого языка или поведения;

– прямой вред пользователю – побуждение к скрытым действиям, таким как формирование мнений, которые кажутся положительными или нейтральными;

– взаимодействия, которые оказывают вредное влияние на поведение, – введение в заблуждение или предоставление ложной информации;

– взаимодействия, которые оказывают вредное влияние на поведение, – манипулирование пользователями и побуждение их делать то, чего они обычно не делают;

– взаимодействия, которые в совокупности наносят вред пользователям – вред, возникающий в процессе отношений с течением времени.

Вред от взаимодействий возникает в результате языкового воздействия (Shelby et al., 2023). Прямой вред пользователю происходит посредством контекстуального языка и языка отношений. Язык может восприниматься как более позитивный (ласковое обращение с использованием пренебрежительных выражений) или менее позитивный (покровительственное отношение к женщинам или пожилым людям) в зависимости от ситуации (Coghlan et al., 2021).

Вторичные негативные последствия возникают в ходе взаимодействий с инфлюэнсерами, то есть лицами, способными влиять на мышление или действия других людей. Таким образом, агентный ИИ может оказывать чрезмерное влияние, вызывая у пользователя несвойственное ему или неадекватное поведение (Alberts, Keeling et al., 2024). Использование социальных сигналов делает системы более интуитивно понятными и привлекательными (Kocielnik et al., 2021). Люди реагируют на социальные сигналы эмоционально, а не рационально, и это может быть использовано для манипулирования их действиями (Alberts, Lyngs et al., 2024; Shamsudhin & Jotterand, 2021).

Взаимодействия, в совокупности причиняющие вред, включают в себя пренебрежительные действия, которые воспринимаются как бестактные

и контролирующие. Коллективный эффект причинения вреда является кумулятивным, например, единичный случай бестактности может быть проигнорирован, но если он повторится, то со временем это будет иметь негативные последствия (Alberts, Keeling et al., 2024).

1.3. Принятие решений с помощью агентного ИИ

Агентные системы искусственного интеллекта требуют беспрецедентной автономии и понимания контекста (Martinez & Kifle, 2024; Mohanarangan et al., 2024). Процесс принятия решений алгоритмом агентного ИИ должен быть поистине революционным, чтобы соответствовать требованиям автономной работы и принятия логичных и последовательных решений в среде, в которой он функционирует. Алгоритм должен принимать решения в режиме реального времени и синтезировать сложные данные и массивы данных (Abuelsaad et al., 2024). Агентный ИИ обладает двумя возможностями, которые выходят за рамки возможностей ИИ-ассистентов. Во-первых, процесс принятия решений осуществляется на разных уровнях, от реагирования на низком уровне до стратегического реагирования на высоком уровне, что требует долгосрочного планирования (Abuelsaad et al., 2024). Вторая возможность – это переход от реактивного к проактивному целенаправленному поведению, которое требует от системы выявления сложных задач и определения необходимых подзадач. Таким образом, для достижения своих целей агентный ИИ требует гибкой архитектуры своего программного обеспечения для управления целями (Martinez & Kifle, 2024). Агентный ИИ должен использовать адаптивный стиль обучения. Это подразумевает освоение различных стилей обучения, а также способность ускорять процесс обучения и активно применять приемы обучения, соответствующие конкретной ситуации. При адаптивной системе обучения система извлекает уроки из прошлого опыта (Abuelsaad et al., 2024).

В деятельности организации агентный ИИ выступает в качестве стороннего исполнителя. Организация, внедряющая систему на ранних этапах, получит преимущество (положение на рынке, инновации, отношения с клиентами, операционная эффективность, уровень обучения, доля рынка). Однако при внедрении ее на поздних этапах компании потенциально потеряют свои конкурентные преимущества (снижение доли рынка и увеличение расходов; задержки при внедрении инноваций; отставание в персонализации услуг; более высокие альтернативные издержки и эксплуатационные расходы; меньше возможностей для раннего обучения; потенциально более высокий барьер для входа на рынок за счет снижения возможностей для тестирования инноваций) (Beulen et al., 2022). Агентный ИИ обладает многими преимуществами, среди которых: положительное влияние на операционные расходы; более высокая эффективность, поскольку ИИ может выполнять задачи автоматически и с большей точностью; масштабируемость без необходимости в дополнительных ресурсах и инвестициях; нацеленность на достижение поставленных целей, тогда как организация может сосредоточиться на своей основной деятельности и оставить второстепенные или менее важные задачи ИИ (Hosseini & Seilani, 2025). Однако использование агентного ИИ имеет и ряд недостатков: зависимость от технологий (чрезмерная зависимость от технологий может привести к сбоям в работе при

использовании искусственного интеллекта); ограниченный диапазон персонализации, тогда как многие задачи требуют тщательной настройки; проблемы конфиденциальности и безопасности, например, передача данных сторонним исполнителям вызывает опасения пользователей по поводу конфиденциальности и безопасности; скрытые расходы, связанные с обучением, развертыванием и внедрением систем.

В будущем приложения агентного ИИ найдут применение во многих отраслях, включая робототехнику и производство[5], системы здравоохранения[6], транспорт и логистику[7], системы управления дорожным движением[8] и финансовые услуги[9]. Одним из новых применений агентного ИИ является динамизация потребностей пациентов, что приведет к созданию персонализированных лекарств (Hasan et al., 2025). При этом агентный ИИ управляет действиями пациентов с хроническими заболеваниями, изучая историю болезни и отправляя напоминания пациентам (Yang, Garcia et al., 2024); для этого необходимо выработать рекомендации по лечению с учетом показателей здоровья. Этот тип агентной системы искусственного интеллекта сможет управлять индивидуальным уходом за пациентами и отслеживать ранние признаки ухудшения состояния здоровья, особенно пожилых лиц (Acharya et al., 2025). Еще одна сфера применения агентного ИИ – автоматическое создание нового контента, ориентированного на широкую аудиторию и отвечающего требованиям на основе установленных критериев. Такое приложение было бы полезно для маркетинговых мероприятий, таких как рассылка персонализированных электронных писем покупателям и потенциальным клиентам. Предприниматели и ученые могут осуществлять быстрый поиск литературы с помощью агентного ИИ, что приведет к появлению новых идей. Агентный ИИ может способствовать открытию, разработке и распространению новых лекарств (Gao et al., 2024).

Исследования с помощью агентного ИИ набирают обороты в области науки о морали и принятия этических решений (Small & Lew, 2021). Важность конфиденциальности и безопасности при работе с чувствительной информацией выдвинула этот тип исследований на передний план. Исследования в области морали направлены на то, чтобы создать этическую основу для автономных систем, чтобы агентные системы ИИ могли выбирать действия с учетом их последствий и ценности. В этом контексте интеграция психологии, этики и философии создает общую цель для систем ИИ, которая является этической. Все агентные системы должны соблюдать этические нормы при принятии решений, но в особенности это касается систем здравоохранения и правопорядка, поскольку решения в этих сферах влияют на общество в целом (Acharya et al., 2025).

---

5 Randieri, C. (2025, January 3). Agentic AI: A New Paradigm In Autonomous Artificial Intelligence. Forbes. https://clck.ru/3NedZf

6 Automation Anywhere. (n.d.). What is agentic AI? Key benefits & features. https://clck.ru/3Nedsx

7 Там же.

8 Randieri, C. (2025, January 3). Agentic AI: A New Paradigm In Autonomous Artificial Intelligence. Forbes. https://clck.ru/3NedZf

9 Там же; Automation Anywhere. (n.d.). What is agentic AI? Key benefits & features. https://clck.ru/3Nedsx

Агентные системы искусственного интеллекта требуют самосознания и метапознания (Langdon et al., 2022). Это может быть достигнуто путем создания систем, понимающих свои действия, способности и ограничения, то есть обладающих самореферентными знаниями. Достичь самосознания в системах искусственного интеллекта можно путем самооценки по следующим направлениям: оптимально ли они выполняли задачи; что можно улучшить; какие действия следует предпринять при возникновении сбоев или низкой производительности. Навыки самоуправления при выполнении задачи и способность определять необходимость ее выполнения позволят агентному ИИ оценивать свои стратегии и процессы обучения, чтобы повысить эффективность принятия решений. Новые достижения в исследованиях самосознания и метапознания приведут к созданию более гибких и сложных агентных систем искусственного интеллекта, что, в свою очередь, повысит производительность и надежность работы в мультисредах (Acharya et al., 2025). Это потребует дальнейших исследований в области создания новых моделей ИИ-агентов, адаптивных моральных норм и контекстуального принятия решений (Lai et al., 2021).

2. Практические рекомендации

ИИ-агенты обладают более высокой эффективностью, чем системы с искусственным интеллектом; поэтому они поднимают и более сложные этические и юридические проблемы. Это усугубляется социальной автономией ИИ-агентов. Пользователь, владелец, разработчик имеют определенную степень контроля над пассивными системами искусственного интеллекта (ИИ-ассистентами), поскольку последние привязаны к определенной позиции и предназначены только для решения определенных проблем или задач.

Владельцы, разработчики и пользователи ИИ-агентов обязаны соблюдать этические принципы при проектировании, внедрении и эксплуатации систем. Технический разработчик и владелец алгоритма должны гарантировать, что агентная система искусственного интеллекта применяется этично и законно. Причина этого состоит в том, что агентное программное обеспечение становится независимым после его выпуска; таким образом, необходимо контролировать его действия и проявления. На ком лежит юридическая ответственность, если система искусственного интеллекта выходит из-под контроля? Вред может быть причинен, например, при получении данных от третьей стороны, поскольку система ИИ обладает определенной способностью принимать решения; некоторые считают ее способной на совершение сознательных действий[10] (Lim et al., 2025). Однако пользователь алгоритма агентного ИИ также несет определенную этическую и юридическую ответственность. Что если пользователь попросит ИИ-агента сделать что-то неэтичное и незаконное, например, передать информацию без соблюдения надлежащей правовой процедуры? Кто будет нести ответственность в этой ситуации – пользователь или разработчик/владелец алгоритма ИИ? Что если отношения между ИИ-агентом и пользователем станут нести опасность, а ИИ-агент выйдет из-под контроля и причинит вред (Alberts, Keeling et al., 2024)? Агентный ИИ может контекстуализировать экологический ландшафт, а значит,

[10] Al-Sibai, N. (2022). OpenAI Chief Scientist Says Advanced AI May Already Be Conscious. https://clck.ru/3Nee2Z

обладает осведомленностью; но позволяет ли это пользователю избежать ответственности? Ситуация несет черты сходства с тем, что происходит в области автономных транспортных средств: стороны пытаются снять с себя вину и ответственность.

Выявленные проблемы не настолько распространены среди ИИ-ассистентов. Переход от пассивных технологических систем к интерактивным вызывает дополнительные опасения не только по поводу юридических и этических последствий, но и по поводу масштабов и возможностей агентных систем с искусственным интеллектом. Агентный ИИ – это будущее направление развития искусственного интеллекта, и его не остановить, учитывая множество преимуществ; однако существуют проблемы, которые необходимо признать и решить для защиты общества и человечества. Уважительное отношение к каждому человеку – это отправная точка для того, чтобы сделать агентный ИИ этически и юридически ответственным. Разработка соответствующих структур должна вестись на основе теории базовых психологических потребностей и с учетом особенностей взаимодействия человека и робота (Li et al., 2025; Hosseini & Seilani, 2025; Korzynski et al., 2025; Kshteri, 2025).

Новые приложения агентного ИИ появляются в сфере здравоохранения, логистики и транспорта, а также финансовых услуг. Однако проблемы безопасности и конфиденциальности остаются актуальными из-за увеличения объема данных и роста автономии при принятии решений с помощью агентного ИИ. Такие системы принимают решения, разбивая сложные задачи на отдельные части. Вопрос в том, насколько надежна архитектура принятия решений и насколько хорошо понимается экологическая экосистема, в которой осуществляется процесс принятия решений. Надежность и точность принимаемых решений основываются на этих факторах и зависят от них. Процесс принятия решений и экологическая экосистема являются отправными точками и основополагающими факторами для получения адекватного результата. Если основополагающие аспекты агентного ИИ недостаточно надежны, алгоритм может выйти из-под контроля и начнет ошибаться. Широкий спектр применений агентного ИИ делает необходимым внедрение защитных механизмов на всех уровнях архитектуры, что, в свою очередь, требует иерархической архитектуры систем агентного ИИ. Однако для тестирования подсистем различных архитектур алгоритма потребуется обратная связь, чтобы можно было выделить или исправить те части, которые работают неэффективно или демонстрируют сомнительные результаты. Потребует ли это избыточности в архитектуре агентного ИИ? Если да, то затраты на приобретение и внедрение агентных систем возрастут. Индикатором верного направления развития может стать наличие логики и осознанности, которые, как предполагается, существуют в системах искусственного интеллекта. Коммуникация агентного ИИ клиентам и потенциальным заказчикам с помощью электронной почты сопряжена с различными бизнес-рисками; таким образом, в системе агентного ИИ необходимы защитные механизмы. Проблемы, влияющие на бизнес, могут нанести ущерб репутации бренда и деловым отношениям. Преимущества применения агентного ИИ в новых и перспективных сферах, таких как разработка лекарств, могут привести к тому, что приложение будет работать без необходимых ограничений и без гарантий надежности существующей архитектуры. Перевешивает ли общественная ценность агентного ИИ преимущества обеспечения строгости нормативно-правовой базы? Должна ли защита агентного ИИ, юридическая и нормативная, основываться в большей степени на методе проб и ошибок или на практическом обучении?

Заключение

Сложные ИИ-агенты (агентный ИИ) обладают множеством преимуществ – от способности работать автономно до способности к рассуждениям; следовательно, они обладают определенным уровнем сознания. Однако существуют риски, которые требуют сбалансированного подхода к внедрению агентного ИИ. Сценарии, уже изученные на примере автономных транспортных средств, применимы и к агентному ИИ, а уроки, извлеченные из опыта производства таких автомобилей, являются хорошей отправной точкой для понимания этических и правовых ситуаций в сфере агентного ИИ. Риски, связанные с ИИ, должны быть сбалансированы соответствующими защитными механизмами, которые не должны препятствовать инновациям в применении ИИ. Это требует развития правовой и этической базы для защиты общества, а также для обеспечения преимуществ ИИ в сфере бизнеса и промышленности. Агентный ИИ переводит процесс принятия решений с интерфейса «человек – машина» на взаимодействие «машина – машина» без необходимости вмешательства человека в принятие решений, но нельзя забывать о рисках. Необходимо установить строгие и в то же время гибкие и надежные защитные механизмы для соблюдения этических и правовых рамок.

Список литературы

Abuelsaad, T., Akkil, D., Dey, P., Jagmohan, A., & Vempaty, A. (2024). Agent-E: From Autonomous Web Navigation to Foundational Design Principles in Agentic Systems. *arXiv preprint arXiv:2407.13032*. https://doi.org/10.48550/arXiv.2407.13032

Acharya, D. B., Kuppan, K., & Ashwin, D. B. (2025). Agentic AI: Autonomous intelligence for complex goals – a comprehensive survey. In *IEEE Access (*vol. 13, pp. 18912–18936). https://doi.org/10.1109/ACCESS.2025.3532853

Alberts, L., Keeling, G., & McCroskery, A. (2024). Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect. *arXiv:2401.09082v2*. https://doi.org/10.48550/arXiv.2401.09082

Alberts, L., Lyngs, U., & Van Kleek, M. (2024). Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–25. https://doi.org/10.1145/3653693

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922

Beulen, E., Plugge, A., & van Hillegersberg. J. (2022). Formal and relational governance of artificial intelligence outsourcing. *Information System E Business Management*, *20*(4), 719–748. https://doi.org/10.1007/s10257-022-00562-7

Coghlan, S., Waycott, J., Lazar, A., & Neves, B. (2021). Dignity, Autonomy, and Style of Company: Dimensions Older Adults Consider for Robot Companions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–25. https://doi.org/10.1145/3449178

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. *Cell*, *187*(22), 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, 113515.

Hasan, S. S., Fury, M. S., Woo, J. J., Kunze, K. N., & Ramkumar, P. N. (2025). Ethical Application of Generative Artificial Intelligence in Medicine. *Arthroscopy: Journal of Arthroscopic Related Surgery*, *41*(4), 874–885. https://doi.org/10.1016/j.arthro.2024.12.011

Hosseini, S., & Seilani, H. (2025). The Role of Agentic AI in Shaping a Smart Future: A Systematic review. *Array*, *26*, 100399. https://doi.org/10.1016/j.array.2025.100399

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI Agents That Matter. *arXiv:2407.01502v1*.

Kocielnik, R., Langevin, R., George, J. S., Akenaga, S., Wang, A., Jones, D. P., Argyle, A., Fockele, C., Anderson, L., Hsieh, D. T., Kabir, Y., Duber, H., Hsieh, G., & Hartzler, A. L. (2021). Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain*. ACM, New York, NY, USA. https://doi.org/10.1145/3469595.3469599

Korzynski, P., Edwards, A., Gupta, M. C., Mazurek, G., & Wirtz, J. (2025). Humanoid robotics and agentic AI: reframing management theories and future research directions. *European Management Journa*l, *43*(4), 548–560. https://doi.org/10.1016/j.emj.2025.06.002

Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. *Telecommunications Policy*, *49*(6), 102976. https://doi.org/10.1016/j.telpol.2025.102976

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. *arXiv:2112.11471*. https://doi.org/10.48550/arXiv.2112.11471

Langdon, A., Botvinick, M., Nakahara, H., Tanaka, K., Matsumoto, M., & Kanai, R. (2022). Meta-learning, social cognition and consciousness in brains and machines. *Neural Network*, *145*, 80–89. https://doi.org/10.1016/j.neunet.2021.10.004

Li, X., Shi, W., Zhang, H., Peng, C., Wu, S., & Tong, W. (2025). The Agentic-AI Core: an AI-Empowered, Mission-Oriented core network for Next-Generation mobile telecommunications. *Engineering*. https://doi.org/10.1016/j.eng.2025.06.027

Lim, S., Schmälzle, R., & Bente, G. (2025). Artificial Social Influence via Human-Embodied AI Agent Interaction in Immersive Virtual Reality (VR): Effects of Similarity-Matching during health conversations. *Computers in Human Behavior Artificial Humans*, *5*, 100172. https://doi.org/10.1016/j.chbah.2025.100172

Martinez, D. R., & Kifle, B. M. (2024). *Artificial Intelligence: A Systems Approach from Architecture Principles to Deployment*. MIT Press eBooks, IEEE Xplore2. https://doi.org/10.7551/mitpress/14806.001.0001

Mohanarangan, S., Karthika, D., Moohambigai, B., & Sangeetha, R. (2024). Unleashing the Power of AI and Machine Learning: Integration Strategies for IoT Systems. *International Journal of Scientific Research in Computer Science and Engineering*, *12*(2), 25–32.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). https://doi.org/10.1145/259963.260288

Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall. Google-Books-ID: CUVeMwAACAAJ.

Shamsudhin, N., & Jotterand, F. (2021). Social Robots and Dark Patterns: Where Does Persuasion End and Deception Begin? In F. Jotterand, & M. Ienca (Eds.), *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues* (pp. 89–110). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-74188-4_7

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *arXiv:2210.05791*. https://doi.org/10.48550/arXiv.2210.05791

Small., C., & Lew, C. (2021). Mindfulness, moral reasoning and responsibility: Towards virtue in ethical decision-making. *Journal of Business Ethics*, *169*(1), 103–117. https://doi.org/10.1007/s10551-019-04272-y

Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., & Press, O. (2024). SWE-AGENT: Agent-Computer Interfaces Enable Automated Software Engineering. *arXiv:2405.15793*. https://doi.org/10.48550/arXiv.2405.15793

Yang, E., Garcia, T., Williams, H., Kumar, B., Ramé, M., Rivera, E., Ma, Y., Amar, J., Catalani, C., & Jia, Y. (2024). From barriers to tactics: A behavioural science-informed agentic workflow for personalized nutrition coaching. *arXiv:2410.14041*. https://doi.org/10.48550/arXiv.2410.14041

Сведения об авторе

Боуэн Гордон– доктор делового администрирования, доцент, школа менеджмента, Университет Англия Рёскин
Адрес: Великобритания, г. Кембридж, CB1 1PT, Ист Роуд
E-mail: gordon.bowen@aru.ac.uk
ORCID ID: https://orcid.org/0009-0007-4082-0336
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56943078600
WoS Researcher ID: https://www.webofscience.com/wos/author/record/65121803
Google Scholar ID: https://scholar.google.com/citations?user=zm_Qgw4AAAAJ

Конфликт интересов

Автор сообщает об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.07.45 / Право и научно-технический прогресс
Специальность ВАК: 5.1.1 / Теоретико-исторические правовые науки

История статьи

Дата поступления – 10 июня 2025 г.
Дата одобрения после рецензирования – 26 июня 2025 г.
Дата принятия к опубликованию – 25 сентября 2025 г.
Дата онлайн-размещения – 30 сентября 2025 г.

Research article
UDC 34:004:340.1:004.8
EDN: https://elibrary.ru/pltfwo
DOI: https://doi.org/10.21202/jdtl.2025.17

Agentic Artificial Intelligence: Legal and Ethical Challenges of Autonomous Systems

Gordon Bowen

Anglia Ruskin University, Cambridge, United Kingdom

Keywords

agentic artificial intelligence,
artificial intelligence,
autonomy,
digital technologies,
ethics,
law,
legal regulation,
liability,
programming,
risk

Abstract

Objective: to identify specific legal and ethical problems of agentic artificial intelligence and develop recommendations for the creation of protective mechanisms to ensure the responsible functioning of autonomous AI systems.

Methods: the research is conceptual in nature and is based on a systematic analysis of scientific literature on the ethics of artificial intelligence, legal regulation of autonomous systems and social interaction of AI agents. The work uses a comparative analysis of various types of AI systems, a study of the potential risks and benefits of agentic artificial intelligence, as well as an interdisciplinary approach that integrates advances in law, ethics, and computer science to form a comprehensive understanding of the issue.

Results: the research has established that agentic artificial intelligence, possessing the decision-making autonomy and ability to social interaction, creates qualitatively new legal and ethical challenges compared to traditional AI assistants. The main categories of potential harm were identified: direct impact on users through overt and covert actions, manipulative influence on behavior, and cumulative harm from prolonged interaction. The author stipulates the need for distributing responsibility between three key actors: the user, the developer and the owner of the agentic artificial intelligence system.

Scientific novelty: for the first time, the research presents a systematic analysis of the ethical aspects of agentic artificial intelligence as a qualitatively new class of autonomous systems that differ from traditional AI assistants in the degree of independence and social interactivity. The author developed a typology of potential risks of social interaction with agent-based intelligent systems and proposes a conceptual model for the distribution of legal and ethical responsibilities in the user-developer-owner triad.

Practical significance: the research forms the theoretical basis for the development of ethical principles and legal norms governing agentic based artificial intelligence in a growing market for autonomous intelligent systems. The findings will be useful for legislators creating a regulatory framework, developers designing protective mechanisms, as well as organizations implementing agentic artificial intelligence systems in various economic fields.

For citation

Bowen, G. (2025). Agentic Artificial Intelligence: Legal and Ethical Challenges of Autonomous Systems. *Journal of Digital Technologies and Law, 3*(3), 431–445. https://doi.org/10.21202/jdtl.2025.17

References

Abuelsaad, T., Akkil, D., Dey, P., Jagmohan, A., & Vempaty, A. (2024). Agent-E: From Autonomous Web Navigation to Foundational Design Principles in Agentic Systems. *arXiv preprint arXiv:2407.13032*. https://doi.org/10.48550/arXiv.2407.13032

Acharya, D. B., Kuppan, K., & Ashwin, D. B. (2025). Agentic AI: Autonomous intelligence for complex goals – a comprehensive survey. In *IEEE Access (*vol. 13, pp. 18912-18936). https://doi.org/10.1109/ACCESS.2025.3532853

Alberts, L., Keeling, G., & McCroskery, A. (2024). Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect. *arXiv:2401.09082v2*. https://doi.org/10.48550/arXiv.2401.09082

Alberts, L., Lyngs, U., & Van Kleek, M. (2024). Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–25. https://doi.org/10.1145/3653693

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922

Beulen, E., Plugge, A., & van Hillegersberg. J. (2022). Formal and relational governance of artificial intelligence outsourcing. *Information System E Business Management*, *20*(4), 719–748. https://doi.org/10.1007/s10257-022-00562-7

Coghlan, S., Waycott, J., Lazar, A., & Neves, B. (2021). Dignity, Autonomy, and Style of Company: Dimensions Older Adults Consider for Robot Companions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–25. https://doi.org/10.1145/3449178

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. *Cell*, *187*(22), 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, 113515.

Hasan, S. S., Fury, M. S., Woo, J. J., Kunze, K. N., & Ramkumar, P. N. (2025). Ethical Application of Generative Artificial Intelligence in Medicine. *Arthroscopy: Journal of Arthroscopic Related Surgery*, *41*(4), 874–885. https://doi.org/10.1016/j.arthro.2024.12.011

Hosseini, S., & Seilani, H. (2025). The Role of Agentic AI in Shaping a Smart Future: A Systematic review. *Array*, *26*, 100399. https://doi.org/10.1016/j.array.2025.100399

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI Agents That Matter. *arXiv:2407.01502v1*.

Kocielnik, R., Langevin, R., George, J. S., Akenaga, S., Wang, A., Jones, D. P., Argyle, A., Fockele, C., Anderson, L., Hsieh, D. T., Kabir, Y., Duber, H., Hsieh, G., & Hartzler, A. L. (2021). Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain*. ACM, New York, NY, USA. https://doi.org/10.1145/3469595.3469599

Korzynski, P., Edwards, A., Gupta, M. C., Mazurek, G., & Wirtz, J. (2025). Humanoid robotics and agentic AI: reframing management theories and future research directions. *European Management Journa*l, *43*(4), 548–560. https://doi.org/10.1016/j.emj.2025.06.002

Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. *Telecommunications Policy*, *49*(6), 102976. https://doi.org/10.1016/j.telpol.2025.102976

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. *arXiv:2112.11471*. https://doi.org/10.48550/arXiv.2112.11471

Langdon, A., Botvinick, M., Nakahara, H., Tanaka, K., Matsumoto, M., & Kanai, R. (2022). Meta-learning, social cognition and consciousness in brains and machines. *Neural Network*, *145*, 80–89. https://doi.org/10.1016/j.neunet.2021.10.004

Li, X., Shi, W., Zhang, H., Peng, C., Wu, S., & Tong, W. (2025). The Agentic-AI Core: an AI-Empowered, Mission-Oriented core network for Next-Generation mobile telecommunications. *Engineering*. https://doi.org/10.1016/j.eng.2025.06.027

Lim, S., Schmälzle, R., & Bente, G. (2025). Artificial Social Influence via Human-Embodied AI Agent Interaction in Immersive Virtual Reality (VR): Effects of Similarity-Matching during health conversations. *Computers in Human Behavior Artificial Humans*, *5*, 100172. https://doi.org/10.1016/j.chbah.2025.100172

Martinez, D. R., & Kifle, B. M. (2024). *Artificial Intelligence: A Systems Approach from Architecture Principles to Deployment*. MIT Press eBooks, IEEE Xplore2. https://doi.org/10.7551/mitpress/14806.001.0001

Mohanarangan, S., Karthika, D., Moohambigai, B., & Sangeetha, R. (2024). Unleashing the Power of AI and Machine Learning: Integration Strategies for IoT Systems. *International Journal of Scientific Research in Computer Science and Engineering*, *12*(2), 25–32.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). https://doi.org/10.1145/259963.260288

Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall. Google-Books-ID: CUVeMwAACAAJ.

Shamsudhin, N., & Jotterand, F. (2021). Social Robots and Dark Patterns: Where Does Persuasion End and Deception Begin? In F. Jotterand, & M. Ienca (Eds.), *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues* (pp. 89–110). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-74188-4_7

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *arXiv:2210.05791*. https://doi.org/10.48550/arXiv.2210.05791

Small., C., & Lew, C. (2021). Mindfulness, moral reasoning and responsibility: Towards virtue in ethical decision-making. *Journal of Business Ethics*, *169*(1), 103–117. https://doi.org/10.1007/s10551-019-04272-y

Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., & Press, O. (2024). SWE-AGENT: Agent-Computer Interfaces Enable Automated Software Engineering. *arXiv:2405.15793*. https://doi.org/10.48550/arXiv.2405.15793

Yang, E., Garcia, T., Williams, H., Kumar, B., Ramé, M., Rivera, E., Ma, Y., Amar, J., Catalani, C., & Jia, Y. (2024). From barriers to tactics: A behavioural science-informed agentic workflow for personalized nutrition coaching. *arXiv:2410.14041*. https://doi.org/10.48550/arXiv.2410.14041

Author information

Gordon Bowen – DBA, Associate Professor, School of Management, Anglia Ruskin University
Address: East Road, CB1 1PT, Cambridge, United Kingdom
E-mail: gordon.bowen@aru.ac.uk
ORCID ID: https://orcid.org/0009-0007-4082-0336
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56943078600
WoS Researcher ID: https://www.webofscience.com/wos/author/record/65121803
Google Scholar ID: https://scholar.google.com/citations?user=zm_Qgw4AAAAJ

Conflict of interest

The author declares no conflict of interest.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – June 10, 2025
Date of approval – June 26, 2025
Date of acceptance – September 25, 2025
Date of online placement – September 30, 2025