УДК 004.83

ИНТЕЛЛЕКТУАЛЬНЫЕ И АВТОНОМНЫЕ СИСТЕМЫ – ЭТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ

П.М. Готовцев (*gotovtsevpm@gmail.com*)
НИЦ Курчатовский институт, Москва

Аннотация. Данная работа представляет собой обзор ряда международных инициатив в области этики интеллектуальных и автономных систем (ИИ/АС). Представлена общая постановка вопроса, в том виде, в котором она отражается в большинстве работ, а именно взаимодействие с человеком ИИ/АС в момент принятия этими системами решения и социальные изменения, связанные с внедрением таких систем. Представлена информация об инициативе Института инженеров в области электротехники и электроники (Institute of Electrical and Electronics Engineers – IEEE), направленной на разработку ряда международных стандартов, относящихся к области этики ИИ/АС. Показаны основные задачи, выделенные в рамках данной инициативы IEEE.

Ключевые слова: правила, интеллектуальные и автономные системы, этика технических систем, человеко-машинное взаимодействия, междисциплинарные исследования, технические стандарты.

Введение

Искусственный интеллект – это словосочетание сегодня звучит в совершенно разном контексте. Его используют как для фантастических сверхразумных машин в кинематографе, так и для продвижения «умных» приложений в обычные смартфоны. Хотя, говоря технически корректно некоторые из таких приложений могут попасть по термин интеллектуальные и автономные системы (ИИ/АС) - именно такая терминология используется применительно к многочисленным «умным» системам, которые так активно входят в нашу жизнь в последние годы. Основной особенностью таких систем, является их способность автономно принимать решения, опираясь на имеющиеся данные. Такой системой может быть, например, автопилот автомобиля, который принимает решения как, с какой скоростью и по какому маршруту ехать. То же можно сказать о программах, которые анализируют наше поведение в интернете – активность в социальных сетях, поисковые запросы, местоположение по

комментариям под фотографиями и т.д. Эти программы, обработав данные дают пользователям советы, конечно, право пользователей следовать или не следовать этим советам, но очень часто получается, что советы оказываются весьма полезными. Таким образом, мы сталкиваемся с ситуацией, когда машины принимают решения, которые непосредственно оказывают влияние на нашу жизнь, и по мере дальнейшего развития технологий мы в праве ожидать что областей применения таких «умных» машин будет становиться все больше. Более того, сегодня наиболее ценные и интересные результаты во многих областях от геномных исследований до интернета вещей получаются там, где проводится анализ больших данных. То есть машины могут принимать решения там, где человек уже не может – при анализе больших данных. Итак можно сказать, что сегодня взаимодействие человека и машин обретает новые особенности, Еще раз обратим внимание на то, что речь идет не о сверхразумной искусственном интеллекте который сейчас активно обсуждается [Bostrom 2014] [Bostrom et al 2011], а о достаточно ограниченных по своим возможностям в сравнении с человеком программных продуктах, которые очень хорошо решают одну определенную задачу и не способны одновременно управлять автомобилем и анализировать медицинские данные что бы поставить диагноз пациенту.

1 Вопросы применения ИИ/АС

Раз мы сказали о новых особенностях взаимодействия человека и машины, то необходимо попробовать сформулировать вопросы, которое в результате этого возникают [Broadbent 2017] [Bostrom et al 2011]. Начнем с некоторых ситуаций:

Автономное транспортное средство везет пассажиров. Водителя нет, роль «таксиста» выполняет ИИ/АС. На дороге встретилась авария, есть пострадавшие, а машины скорой помощи еще нет. Что делать в этой ситуации – объехать место аварии, остановится и дать возможность пассажирам убедиться, что скорая в пути? А если среди пассажиров врач, который может помочь?

Домашний робот, осуществляет простейшую уборку пола. Для совершенствования своего обучаемого программного обеспечения передает все данные компании-разработчику. Только вот этот незаметный домашний робот, по сути, передает данные о частной жизни пользователя, организации, которая никак не обязана и не способна заботиться об их сохранности.

Программа, анализирующая поведение пользователя в сети, может получить о пользователе те данные, которые он предпочел бы скрыть.

Кроме того, возникают многочисленные вопросы, связанные с взаимодействием человек – машина, включая выстраивание диалога с программой-советчиком, степень доверия к ИИ/АС, взаимодействия робототехнических систем и человека и т.д. [Grinbaum et al 2017] [Wiener 1960] [Lemagnan et al 2017].

Другим примером являются интеллектуальные программы советчики [Noothigattu et al. 2017] [Gotovtsev et al 2008] [Flores-Loredo et al 2005]. Использование таких программ поднимает целый ряд этических вопросов, среди которых можно отметить следующие:

1. Афилированность программ-советчиков;
2. Влияние таких программ на способность пользователей самостоятельно принимать решение;
3. Может ли, например, программа, анализирующая состояние потенциального заемщика, учесть особые факторы, связанные со сложной жизненной ситуацией?
4. Программа-советчик и опыт оператора – взаимодополнение или противопоставление?

Немаловажным является вопрос анализа метаданных, уже упомянутый выше, а также производимые интеллектуальными системами продукты, как в области искусства [Knight 2011] [Elgammal et al 2017], так и товаров потребления [Yoon et al 2017].

Таким образом, можно сказать, что вопросы этики ИИ/АС возникают во многих областях, где такие системы находят сейчас активное применение. От банковских систем и программ-советчиков, до беспилотных мобильных систем (автомобили и дроны-доставщики) и домашних роботов.

2 Инициатива IEEE

Вопросов, вроде указанных в предыдущем разделе, сегодня возникает все больше, и все они складываются в то направление исследований, которое сегодня получило название этика ИИ/АС [Havens 2016] [EAD 2019]. Исследования в этой области привели к появлению ряда инициатив по исследованию в данной области [Grinbaum et al 2017] [Bostrom et al 2011]. Среди таких инициатив следует выделить глобальную инициативу Института инженеров в области электротехники и электроники (Institute of Electrical and Electronics Engineers – IEEE, https://www.ieee.org/), в рамках которой начата работа по созданию нормативно-технических документов, которые заложили бы основу этичного поведения в разрабатываемые системы с ИИ/АС. Первым шагом в этой работе является создание рекомендаций, направляющих разработчиков ИИ/АС на этически обоснованные решения при разработке и применении систем ИИ/АС [EAD 2019]. Эти рекомендации распространяются как на ситуацию, в которой

система с ИИ/АС или автономная система принимает решение в какой-либо этической ситуации, так и на ситуацию, в которой применение ИИ/АС или автономных систем приводит к каким-либо социальным последствиям в виде сокращения определенного персонала. Таким образом, можно выделить два направления исследований [Havens 2016]:

- непосредственно этику ИИ/АС, как этические проблемы, вызванные работой машин, самостоятельно принимающих решения и так или иначе взаимодействующих с человеком;
- этические проблемы, связанные с массовым внедрением ИИ/АС, которые могут привести к сокращениям персонала, занятого рутинными работами и другим социальным вызовам.

В этой статье мы не будем акцентировать внимание на втором вопросе, так как сегодня еще нет достаточно однозначного мнения о том, как и насколько повлияет на рабочие места эффект от массового внедрения ИИ/АС, как это скажется на различных слоях общества. Будет ли только эффект сокращения занятости или это приведет еще и к росту экономики и появлению, например, безусловного дохода? Кроме того, многие аналитические работы концентрируются на сокращении числа рабочих мест в одних областях, но не учитывают потенциала роста других. Так, например, не учитывается рост биотехнологической индустрии, сельского хозяйства, авиаперевозок, где в ряде развитых стран отмечается даже недостаток рабочей силы. В общем это благодарное поле для исследований ученым из многих областей наук, которое стало актуальным со времен промышленной революции и с начала массового внедрения автоматизации [Wiener 1960] [Markoff 2015].

Очевидно, что в конечном итоге все исследования в этой области ведут к созданию нормативных документов, чем и занимается IEEE (https://ethicsinaction.ieee.org/). При этом, что этическое регулирование ИИ/АС и автономных систем не ведет к ограничениям или запрету в развитии таких технологий. Например, система «умный дом» может вызвать помощь, если определит, что человек в доме, например, потерял сознание. Эта система принимает решение, вызывая помощь, и тем самым, она может спасти жизнь человеку. Ограничивать развитие таких систем, значит лишать шанса многих людей в будущем.

В 2017 году IEEE запустило проект под названием «Глобальная инициатива по этике интеллектуальных и автономных систем». В рамках этой инициативы были собраны мнения более чем 2000 специалистов, как ученых из разных областей наук, так и инженеров. На основе этой информации был разработан концептуальный документ, цель которого поднять вопросы, связанные с использованием ИИ/АС и их взаимодействия с человеком, а также указать направления для исследований в этой области [EAD 2019]. Охватываемые документов вопросы включают в себя:

- Обсуждение возможности создания методологии для разработчиков ИИ/АС, учитывающей этичность поведения таких систем в будущем;
- Вопросы связанные с валидацией ИИ/АС в части их взаимодействии с человеком;
- Вопросы связанные с базовыми подходами к разработке ИИ/АС с учетом этически обусловленного взаимодействия с человеком. Причем в данном случае авторы делают акцент на особенности устоявшихся моральных норм в различных сообществах и различные виды ценностей, не ограничиваясь только базовыми ценностями западной цивилизации, но и учитывая особенности восточных и африканских мировоззрений;
- Вопросы подготовки специалистов, связанные с этикой ИИ/АС;
- Юридические вопросы, в частности юридического статуса ИИ/АС, ответственности разработчика и ответственности пользователя;
- Потенциальные вызовы применения ИИ/АС в различных областях.

Опираясь на данный документ [EAD 2019], IEEE создает целую группу стандартов, посвященных этическим аспектам ИИ/АС. На сегодня эта группа включает в себя 13 различных стандартов, получивших проектные номера, начиная с P7000. Они охватывают ряд аспектов связанных с применением ИИ/АС при обработке персональных данных, в том числе различных социальных и возрастных групп, взаимодействию машин и людей (роботы-сиделки, распознавание лиц и т.д.), фильтрации фейковых новостей в информационном потоке, онтологиям для ИИ/АС и самое важное вопросам валидации ИИ/АС и этическим аспектам их разработки (https://ethicsinaction.ieee.org/). Следует отметить, что данные стандарты являются нормативно техническими документами, то есть должны стандартизировать то, как проводить определенные работы по разработке системы или то каким требованиям должна соответствовать в ходе своей эксплуатации ИИ/АС. Таким образом, возможно впервые разработкой технических стандартов занимается междисциплинарная группа ученых из самых разных областей знаний.

3 Исторические предпосылки инициативы IEEE

Вопрос этики ИИ/АС возник не сегодня, Норберт Винер затрагивает вопросы этики ИИ в своей статье [Wiener 1960] и в дополнительных главах второго издания своей «Кибернетики» [Wiener 1965]. Основная мысль, которую он постулирует в этих работах, заключается в том, что машину могут быть опасны для человека и непредсказуемы. При этом он отмечает, что даже понимая в деталях как работает машина, оператор может не успеть понять, что ее рассуждения идут к негативному сценарию или даже не

успеть понять, что машина уже «приняла решение» и работает над осуществлением этого сценария. Уже к моменту написания статьи [Wiener 1960] разница в скорости обработки информации требующей вычислений между человеком и современными на тот момент вычислительными машинами была значительной. Именно эта разница, помноженная на неполную конкретизацию человеком своих желаний и может привести по мнению автора к негативным последствиям. Как видно из вышесказанного, поставленные Норбертом Винером вопросы и сегодня остаются актуальными, при этом зачастую сегодня оператор и разработчики не имеют информации о том, что происходит в данный момент времени в созданных ими интеллектуальных агентах. Уже классическим примером являются масштабные искусственные нейронные сети (ИНС) глубокого обучения, прописать детально работу некоторых из них после обучения не в состоянии даже сами разработчики. Тем не менее этот крайне эффективный метод сегодня получает широчайшее распространении при анализе данных, включая поиск метаданных, а также разработаны алгоритмы, позволяющее таким ИНС иметь «ячейки памяти» для сохранения промежуточных состояний [Graves et al 2016] [Kirkpatrick et al 2017]. В то время как для взаимодействия ИНС с людьми при решении совместных задач уже приходится разрабатывать новые алгоритмы [Crandall et al. 2017].

4 Заключение

Тем не менее, многие читающие сейчас эти строки разработчики ИИ/АС наверняка думают что-то вроде – ну вот сейчас придут гуманитарии и начнут нас своей этикой ограничивать. Во-первых, такие исследования удел не только философов и культурологов, но и инженерного сообщества. Так как перед разработчиками ИИ/АС возникает задача создать алгоритмы способные в области своей деятельности вести себя с учетом тех этических основ, которые будут выработаны научным сообществом. Решение этих задач может стать еще одним из двигателей прогресса в области ИИ/АС, помогая построить мост между формальными математическими системами и гуманитарными понятиями как этика, нравственность и мораль [Карпов и др 2018].

В завершение следует отметить, что решение задач в области этики ИИ, очевидно, требует междисциплинарного подхода. Этот подход должен выражаться не только в исследованиях с участием ученых из разных областей науки, но и появлением в учебных курсах специалистов в области ИИ вопросов этики, и наоборот, предметов, рассказывающих о современном уровне технологий для тех студентов-гуманитариев, кто интересуется технической этикой в целом.

Список литературы

[Карпов и др 2018] В.Э. Карпов, П.М. Готовцев, Г.В. Ройзензон. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018, №2(87) С.84-105

[Bostrom 2014] N. Bostrom, Superintelligence: Paths, Dangers, Strategies. 2014.

[Bostrom et al 2011] N. Bostrom and E. Yudkowsky, The Ethics of Artificial Intelligence // Cambridge Handb. Artif. Intell., pp. 1–20, 2011.

[Broadbent 2017] E. Broadbent, Interactions With Robots: The Truths We Reveal About Ourselves // Annu. Rev. Psychol., vol. 68, no. 1, pp. 627–652, Jan. 2017.

[Crandall et al. 2017] J. W. Crandall et al., "Cooperating with Machines," Mar. 2017.

[EAD 2019] Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. 2019.

[Elgammal et al 2017] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "CAN: Creative Adversarial Networks, Generating 'Art' by Learning About Styles and Deviating from Style Norms," 2017.

[Flores-Loredo et al 2005] Z. Flores-Loredo, P. H. Ibarguengoytia, and E. F. Morales, On Line Diagnosis of Gas Turbines using Probabilistic and Qualitative Reasoning // Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems, 2005, vol. 2005, pp. 297–301.

[Gotovtsev et al 2008] P. M. Gotovtsev, V. N. Voronov, and D. S. Smetanin, Analysis of the coolant condition with the help of artificial neural networks, // Therm. Eng., vol. 55, no. 7, pp. 552–557, Jul. 2008.

[Grinbaum et al 2017] B. A. Grinbaum, R. Chatila, L. Devillers, J. Ganascia, C. Tessier, and M. Dauchet, Ethics in Robotics Research // IEEE Robot. Autom. Mag., vol. 24, no. 3, pp. 139–145, 2017.

[Graves et al 2016] A. Graves et al., Hybrid computing using a neural network with dynamic external memory // Nature, 2016.

[Havens 2016] J. Havens, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines. 2016.

[Kirkpatrick et al 2017] J. Kirkpatrick et al., Overcoming catastrophic forgetting in neural networks // Proc. Natl. Acad. Sci. U. S. A., p. 201611835, Mar. 2017.

[Knight 2011] H. Knight, Eight Lessons learned about Non-verbal Interactions through Robot Theater // ICSR: International Conference on Social Robotics, 2011, pp. 42–51.

[Lemagnan 2017] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, Artificial cognition for social human–robot interaction: An implementation // Artif. Intell., vol. 247, pp. 45–69, Jun. 2017.

[Markoff 2015] J. Markoff, Machines of Loving Grace. Ecco Pr, 2015.

[Noothigattu et al. 2017] R. Noothigattu et al., "A Voting-Based System for Ethical Decision Making," 2017.

[Wiener 1960] N. Wiener, Some Moral and Technical Consequences of Automation // Science (80-.)., vol. 131, no. 3410, pp. 1355–1358, 1960.

[Wiener 1965] N. Wiener, Cybernetics : or, Control and communication in the animal and the machine. M.I.T. Press, 1965.

[Yoon et al 2017] H.-S. Yoon et al., CAD/CAM for scalable nanomanufacturing: A network-based system for hybrid 3D printing // Microsystems Nanoeng., vol. 3, p. 17072, Sep. 2017.